Fairness Is More Than Algorithms: Racial Disparities in Time-to-Recidivism

Jessy Xinyi Han
Massachusetts Institute of Technology
xyhan@mit.edu

Kristjan Greenewald
IBM Research
Kristjan.H.Greenewald@ibm.com

Devavrat Shah
Massachusetts Institute of Technology
devavrat@mit.edu

Abstract

Racial disparities in recidivism remain a persistent issue within the criminal justice system, increasingly exacerbated by the adoption of algorithmic risk assessment tools for decision making. Past works have primarily focused on understanding the bias induced by algorithmic tools, viewing recidivism as a binary outcome—i.e., reoffending or not. Limited attention has been given to the role of non-algorithmic factors (including socioeconomic ones) in driving the racial disparities in recidivism from a systemic perspective. Towards that end, this work presents a multi-stage causal framework to investigate the advent and extent of racial disparities by considering the time-to-recidivism rather than a simple binary outcome. The framework captures the interactions between races, the risk assessment algorithm, and contextual factors in general. This work introduces the notion of counterfactual racial disparity and offers a formal test using survival analysis that can be conducted with observational data to understand whether potential differences in recidivism rates among racial groups arise from algorithmic bias, contextual factors, or their interplay. In particular, it is formally established that if sufficient statistical evidence for differences in recidivism across racial groups is observed, it would support rejecting the null hypothesis that non-algorithmic factors (including socioeconomic ones) do not affect recidivism. An empirical study applying this framework to the COMPAS dataset reveals that short-term recidivism patterns do not exhibit racial disparities when controlling for risk scores. However, statistically significant disparities emerge with a longer follow-up period, particularly for low-risk groups. This suggests that factors beyond the algorithmic scores–possibly including structural disparities in housing, employment, and social support–may accumulate and exacerbate recidivism risks over time. Indeed, the use of survival analysis enables such nuanced analysis. This empirical analysis underscores the need for holistic policy interventions extending beyond algorithmic improvements to address the broader influences on recidivism trajectories.

1 Introduction

With millions of formerly incarcerated people returning to prisons each year, recidivism—the cycle of re-offending following release from incarceration—remains a pressing challenge worldwide. In the United States, recidivism is closely entwined with racial, economic, and social inequalities that permeate the criminal justice system. Minority groups often face disparate treatment across various stages of the process, including 911 call for service [11], policing [10], court sentencing [16],

NeurIPS 2025 Workshop on MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

probation and parole decisions [13, 12], and re-entry support [19]. Thus a close-up examination of the pathways and extent of such disparities must precede any effective reforms for a fair system.

Amid these systemic challenges, algorithmic risk assessment tools such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [9] added another layer of complexity. Designed to reduce human biases in bail, parole, and sentencing decisions, these tools have received continual scrutiny and criticism [3, 2, 6]. ProPublica's influential report on machine bias [1], which analyzed COMPAS [14], suggested racial disparities in predictive accuracy: African American individuals who did not recidivate within two years were disproportionately labeled higher risk compared to Caucasian counterparts. Although ProPublica's emphasis on *equalized odds* highlighted static bias in algorithmic decision-making, it does not account for broader structural contexts or how bias propagates through the criminal justice system over time. More details in Appendix A.

Contributions. The primary contribution of this work is to systematically address the question: to what extent do racial disparities in recidivism, often attributed to algorithmic bias, actually stem from broader contextual factors? The key challenge lies in disentangling the interactions between algorithmic decisions, race, and additional contextual factors over time.

We propose a multi-stage causal framework that captures the trajectory from arrest to re-offense or return to custody. This allows us to examine both direct and indirect pathways of disparities. While contextual factors such as housing or employment are often unobserved, we assume algorithmic risk assessments serve as potentially biased yet informative proxies for demographic and prior crime histories.[1] In other words, given fixed contexts and algorithmic decisions, race itself does not determine time-to-recidivism.

Building on this framework, we introduce the notion of counterfactual racial (dis-)parity, a fairness criterion that examines whether individuals of different races—but otherwise identical—exhibit equivalent time-to-recidivism patterns. We move beyond static fairness measures that treat outcomes as binary predictions to consider, through survival analysis, the dynamic nature of recidivism affected by structural inequality over time.

To assess whether disparities are driven by algorithmic predictions only, or by additional factors, we arrive at Theorem 1 and Lemma 1 to formulate a test around the recidivism curves of different racial groups with the same risk score group. The challenge is that true time-to-recidivism is masked by censoring, as individuals may not re-offend before returning to custody for non-criminal reasons. We therefore leverage the log-rank test from survival analysis, which accounts for censoring, to provide a formal empirical test. If statistical evidence supports that recidivism curves differ by race, we reject the null hypothesis that additional contexts do not directly affect time-to-recidivism.

We analyze the COMPAS dataset curated by ProPublica. Within a short-term follow-up of up to 7 months, we do not find sufficient evidence that recidivism patterns differ across racial groups. However, disparities become significant with follow-up periods exceeding seven months, particularly among individuals categorized as low risk, thereby rejecting the hypothesis that algorithmic bias alone explains the observations. A plausible explanation is that structural inequalities in socioeconomic conditions, including access to housing, employment, and social support, exert a cumulative influence over time, extending beyond algorithmic predictions. We thus advocate for comprehensive policy interventions addressing these broader determinants.

Organization. The paper is organized as follows. In Section 2, we lay out the theoretic foundation of our multi-stage causal framework and introduce a data-driven test for contextual effects. In Section 3, we apply the framework to the COMPAS dataset and conclude with a discussion of potential contextual factors and policy reforms to combat systemic racism.

2 Unpacking Racial (Dis)parities in Recidivism: A Causal Framework

Recidivism is a complex and systemic issue, influenced by social, economic, and institutional factors. To understand the advent and extent of racial disparities among individuals with comparable risk profiles, we propose a multi-stage causal framework that captures the full trajectory—from arrest to

[1] One notion of bias can be taken from *disparate treatment*, where risk assessment algorithms directly use race or protected attributes as inputs.

reoffense or return to custody. This framework makes explicit how perceived race, algorithmic risk assessments, and additional contextual factors may interact to shape outcomes over time.

Framework. We consider a cohort of arrested individuals subject to the COMPAS risk assessment tool. These individuals undergo pre-trial and sentencing decisions, and once released, face three possible outcomes: successful reintegration, reoffense and rearrest, or return to custody for non-criminal violations. We define recidivism as the target event, measured as time from release to rearrest. Return to custody for non-criminal reasons is a censoring event that masks the potential occurrence of recidivism.

Let $D \in \{\text{majority}, \text{minority}\}$ denote race, $M \in \{\text{low}, \text{medium}, \text{high}\}$ the risk category, τ time to recidivism, τ' time to custody, and $T = \min\{\tau, \tau'\}$ the observed time. Contextual factors U (e.g., socioeconomic conditions) may influence several of these variables. A key assumption is that, given context U and a fully informative but possibly biased risk score M, race does not directly make someone reoffend sooner or later. We summarize this in the causal DAG in Figure 1 and defer details to Appendix B.

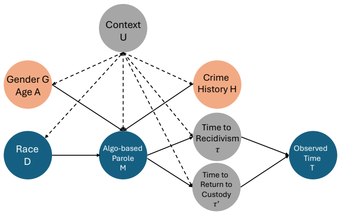

Figure 1: A causal DAG corresponding to the multi-stage recidivism process.

Given our causal framework, we now formalize the notion of racial parity by examining how the *intervention* of race D affects time-to-recidivism under varying contexts.

Definition 1 (Counterfactual Racial Parity)**.** *The system exhibits* counterfactual racial parity *if* $\forall t > 0, m \in \{\textit{low, medium, high}\}$, $\mathbb{P}^{do(D=\textit{majority})}[\tau > t \mid M = m] = \mathbb{P}^{do(D=\textit{minority})}[\tau > t \mid M = m]$.

This definition can be understood as a thought experiment: would two individuals identical in every respect except perceived races experience the same recidivism trajectory under same algorithmic decisions? Unlike static fairness notions, this highlights *when* disparities emerge over time.

Hypothesis and Empirical Test. Having established the multi-stage causal framework and ideal goal of achieving counterfactual racial parity, we face a key challenge: the unobservability of contextual factors U, which may create spurious associations between race D and time-to-recidivism τ and make it difficult to directly assess whether counterfactual racial parity holds. To address this, we examine the role of context through hypothesis testing. Specifically, we do hypothesis testing of a necessary condition under the null hypothesis – absence of such spurious association – to verify if the additional contextual effects indeed exist using real-world data. This leads us to first formulate the following hypothesis about the structural role of context and then offer a formal empirical test.

Hypothesis 1 (Structural Hypothesis)**.** *H_0: context U does not directly affect τ or τ'.* *H_1: context U directly affects τ or τ'.*

Under H_0, we obtain the following implication:

Theorem 1. *If H_0 holds, then for all $t > 0$ and m, $\mathbb{P}^{do(D=d)}[\tau > t \mid M = m] = \mathbb{P}[\tau > t \mid M = m]$.*

Corollary 1. *If H_0 holds, counterfactual racial parity follows automatically within each risk group.*

Thus, observing different survival curves across races within the same risk category allows us to reject H_0 and infer direct contextual effects beyond what is captured by the algorithm. This leads to the following test.

Empirical Test 1. *Let $S_d(t|m) := \mathbb{P}[\tau > t|D = d, M = m]$ $\forall d \in \{\textit{majority}, \textit{minority}\}$. Then $\forall m \in \{\textit{low, medium, high}\}$,*

$$\hat{H}_0(m) : \{S_{\textit{majority}}(t|m) = S_{\textit{minority}}(t|m) \mid t > 0\} \text{ VS. } \hat{H}_1(m) : \{S_{\textit{majority}}(t|m) \neq S_{\textit{minority}}(t|m) \mid t > 0.\}$$

We use the log-rank test to accomplish this empirical test. Details on the test statistic, assumptions, and implementation are deferred to Appendix C.

3 Empirics

Having developed our causal framework and test, we now apply them to the ProPublica COMPAS data. It contains demographic and criminal history information for roughly 10,000 defendants in

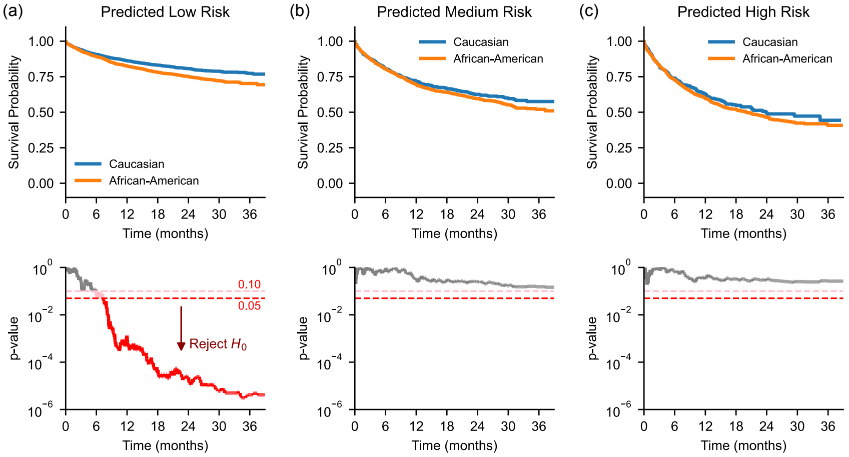

Figure 2: Survival analysis of recidivism across racial groups and COMPAS risk groups. (a) survival curves for Caucasians, (b) survival curves for African-Americans, and (c) corresponding p-values from log-rank tests over time. Gray ($p > 0.1$) indicates insufficient evidence of racial differences, light pink ($0.05 < p \leq 0.1$) marginal differences, and red ($p \leq 0.05$) significant differences.

Broward County, Florida (2013–2014 assessments with follow-up through 2016). Following prior work, we focus on Caucasian (*majority*) and African-American (*minority*) groups, and categorize COMPAS scores into low, medium, and high risk. Returns to custody are censoring events. Data preprocessing details and robustness checks are provided in the Appendix D.

Results. Our analysis reveals distinct temporal patterns across risk groups. For individuals in the *medium* and *high* risk categories, survival curves for majority and minority defendants remain similar, and log-rank tests show no statistically significant differences ($p > 0.1$). This suggests that within these groups, algorithmic risk scores largely explain observed outcomes. In contrast, for those in the *low* risk group, survival curves initially overlap but begin to diverge after approximately seven months. At this point, the log-rank test rejects $\hat{H}_0$, showing that minority defendants experience faster declines in no-recidivism probability than majority defendants assigned the same risk score group. For robustness, we randomly shift 10% of the majority race's score groups higher and 10% of the minorty race's score groups lower and reconduct the empirical test for the perturbed low risk score group and find similar significant results as in Figure 4.

Discussion. These results suggest that algorithmic bias alone cannot explain disparities in the long run. In the short term, outcomes appear comparable across races within the same risk category. But over time, contextual factors not captured by the risk score, such as housing, employment, or social support, may exert cumulative influence, disproportionately affecting minority defendants. The fact that this divergence is concentrated in the low-risk group is particularly troubling, as these individuals are otherwise assessed as having the highest potential for successful reintegration. In line with our framework, we conclude that disparities emerge over time due to contextual influences beyond algorithmic scores. This points toward the need for policy interventions that extend beyond improving algorithms to addressing the broader socioeconomic conditions that shape recidivism trajectories. Extended figures (including violent recidivism results, per-score analyses) are in Appendix D.

4 Conclusion

Time-aware fairness reveals dynamics hidden by binary metrics. Our framework converts a causal question *are there disparities beyond algorithms?* into a survival-curve test usable with observational data. On COMPAS, short-term gaps are limited within risk groups, yet low-risk individuals show significant long-run divergence across races. Policy should pair algorithmic scrutiny with structural supports (housing, employment, supervision quality). Extensions to credit, healthcare triage, and supervision settings are immediate.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

[2] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 62–76. PMLR, 23–24 Feb 2018.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[4] Marco Castillo, Sera Linardi, and Ragan Petrie. Recidivism and Barriers to Reintegration: A Field Experiment Encouraging Use of Reentry Support, 2024.

[5] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[7] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html#document/p32/a310125, July 2016.

[8] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. 80:38, 2016.

[9] Northpointe Institute for Public Management. Compas [computer software], 1996.

[10] Roland G. Fryer. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*, 127(3):1210–1261, 2019.

[11] Jessy Xinyi Han, Andrew Cesare Miller, S. Craig Watkins, Christopher Winship, Fotini Christia, and Devavrat Shah. A causal framework to evaluate racial bias in law enforcement systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):562–572, 2024.

[12] Beth M. Huebner and Timothy S. Bynum. The Role of Race and Ethnicity in Parole Decisions. *Criminology*, 46(4):907–938, 2008.

[13] Anat Kimchi. Investigating the Assignment of Probation Conditions: Heterogeneity and the Role of Race and Ethnicity. *Journal of Quantitative Criminology*, 35(4):715–745, 2019.

[14] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, May 2016.

[15] Drago Plečko and Elias Bareinboim. Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

[16] M. Marit Rehavi and Sonja B. Starr. Racial Disparity in Federal Criminal Sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

[17] Cynthia Rudin, Caroline Wang, and Beau Coker. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1), mar 31 2020. https://hdsr.mitpress.mit.edu/pub/7z10o269.

[18] Maya Sen and Omar Wasow. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Review of Political Science*, 19(1):499–522, 2016. _eprint: https://doi.org/10.1146/annurev-polisci-032015-010015.

[19] Bruce Western and Catherine Sirois. Racialized Re-entry: Labor Market Inequality After Incarceration. *Social Forces*, 97(4):1517–1542, 2019.

A More Details on COMPAS Algorithmic Bias Discussion

Rigorously speaking, their main findings only test how different is *a variant* of the two races' actual false positive rate and false negative rate.

Mathematically, the comparison of the actual false positive rate and false negative rate is defined as

$$\mathbb{P}(M \in \{\text{medium, high}\} | D = \text{majority}, \tau > 2) \gtrless \mathbb{P}(M \in \{\text{medium, high}\} | D = \text{minority}, \tau > 2)$$
$$\mathbb{P}(M \in \{\text{low}\} | D = \text{majority}, \tau \leq 2) \gtrless \mathbb{P}(M \in \{\text{low}\} | D = \text{minority}, \tau \leq 2)$$

where $M \in \{\text{low, medium, high}\}$ denotes the algorithmic risk assessment decision, $D \in \{\text{majority, minority}\}$ denotes the race, and τ denotes the actual time to recidivism. However, the true time to recidivism is often masked by the time to return to custody for non-criminal violations, meaning if returning to custody happens first, then we only observe the minimum of the two, time to return to custody, instead of the target time to recidivism. This is referred to as the right-censoring problem in survival analysis, requiring more careful time-to-event examination.

This work has also sparked intense debate over using *equalized odds* in criminal justice settings [17, 8]. [7] and subsequent responses defended COMPAS's *predictive parity*, i.e., $\mathbb{P}(\tau \leq 2 | D = \text{majority}, M \in \{\text{low}\}) \simeq \mathbb{P}(\tau \leq 2 | D = \text{minority}, M \in \{\text{low}\})$, arguing that its design and operational goals inherently prioritized predictive consistency and accuracy, not necessarily equity. In fact, as shown by [5], so long as the base rate of the two populations differs, i.e., $\mathbb{P}(\tau \leq 2 | D = \text{majority}) \neq \mathbb{P}(\tau \leq 2 | D = \text{minority})$, *equalized odds* and *predictive parity* cannot hold simultaneously for any non-trivial not-perfect classifier.

B More Details on Framework

Arrested Individual. Let $D \in \{\text{majority, minority}\}$ denote the race of the arrested individual.[2]

Algorithm-based Decision. The criminal justice system uses algorithmic risk scores to inform decisions about bail, parole, and probation, potentially shaping an individual's post-release trajectory. We use $M \in \{\text{low, medium, high}\}$, the assigned risk score category, as a proxy for the algorithm-based criminal justice system decisions. We assume such risk assessment scores are fully informative (but likely biased) characterization of demographic features like race D, age, gender, crime history and other contextual background information.

Recidivism or Returning to Custody. Upon release, the individuals are followed up till they re-offend and are rearrested, they return to custody for non-criminal violations or the follow-up period ends, whichever comes first. Specifically, recidivism is the target event and returning to custody is the censoring event. We denote by τ the true time to recidivism, potentially unobserved in certain cases if masked by time to return to custody τ'. T is the observed time, determined entirely by τ and τ', i.e. $T = \min\{\tau, \tau'\}$.

Context. Socioeconomic conditions and other contextual factors U may influence multiple variables in our framework: the individual's race D, demographic characteristics, algorithm-based criminal justice system decision M, time to recidivism τ, and time to return to custody τ'. However, the context information is generally exogenous. Note that we adopt the dashed bidirectional arrow notation ($\leftarrow -- \rightarrow$) as in [15] between protected attributes D and the context U to denote the associational relationship, instead of a causal one.

Causal Mechanism. The causal relationship within this framework is governed by the interactions between context U, race D, algorithm-based criminal justice system decision M, and timing variables τ, τ', and T. As noted, the context U may influence D, M, τ, and τ'. The algorithm-based decision M, capturing race D, gender, age, and crime history, may affect the observed time T through the potential time-to-recidivism τ and time-to-custody τ'. The underlying model is represented in a causal Directed Acyclic Graph (DAG) in Figure 1. A key assumption encoded in this causal mechanism is

[2] Although extensive literature underscores the socially constructed nature of racial categories [18], the data constraints in large-scale recidivism studies make it challenging to adopt a fully constructivist approach. As a consequence, we follow the convention of much of the causal criminology literature, which often employs a non-constructivist perspective to align with existing studies and ensure comparability.

that there is *no direct arrow* between D and τ, τ'. In words, this encodes *given the societal* context *and a fully informative (but likely biased) proxy algorithm-based criminal justice system decision, race does not make someone recidivate sooner or later.*

B.1 Proof of Theorem 1

Proof. Under H_0, we remove the edges $U \to \tau$ and $U \to \tau'$ from Fig. 1. Thus, D and τ are d-separated by M, i.e. $D \perp \tau | M$.

Based on the modified DAG and do calculus, we have $\forall d \in \{\text{majority}, \text{minority}\}, t > 0, m \in \{\text{low, medium, high}\}$

$$\begin{aligned}
&\mathbb{P}^{\text{do}(D=d)}[\tau > t | M = m] \\
=&\sum_u \mathbb{P}^{\text{do}(D=d)}[\tau > t, U = u, D = d | M = m] \\
=&\sum_u \mathbb{P}^{\text{do}(D=d)}[\tau > t, U = u | D = d, M = m] \\
&\quad \cdot \mathbb{P}^{\text{do}(D=d)}[D = d | M = m] && (1) \\
=&\sum_u \mathbb{P}[\tau > t, U = u | D = d, M = m] && (2) \\
=&\mathbb{P}[\tau > t | D = d, M = m] && (3) \\
=&\mathbb{P}[\tau > t | M = m] && (4)
\end{aligned}$$

where (2) is obtained from (1) due to $\mathbb{P}^{\text{do}(D=d)}[D = d | M = m] = 1$ and (4) is obtained from (3) due to $D \perp \tau | M$. □

B.2 Proof of Corollary 1

Proof. Since Theorem 1 holds for $\forall d \in \{\text{majority}, \text{minority}\}$, we have $\mathbb{P}^{\text{do}(D=\text{majority})}[\tau > t | M = m] = \mathbb{P}[\tau > t | M = m] = \mathbb{P}^{\text{do}(D=\text{minority})}[\tau > t | M = m]$. □

Key Implication. Theorem 1 shows that, under the system structures encoded in the causal DAG and null hypothesis H_0, the causal quantity no-recidivism probability $\mathbb{P}^{\text{do}(D=d)}[\tau > t | M = m]$—which reflects an intervention on race—can be expressed directly in terms of a statistical quantity $\mathbb{P}[\tau > t | D = d, M = m]$. This quantity can be further reduced to $\mathbb{P}[\tau > t | D = d]$ since D is independent of τ conditioning on M under H_0. The intuition behind this and Corollary 1 is that since the algorithm-based criminal justice system decision is fully informative, when risk scores fully explain the disparities and additional contextual factors have no direct effect on recidivism timing, controlling for algorithmic decisions alone should ensure counterfactual racial parity.

Moreover, the contrapositive argument of Theorem 1 leads to a practical test: if we observe different recidivism patterns across racial groups within the same algorithmic decision category, we can reject H_0, which implies the sufficiency of algorithmic scores alone to explain the observed disparities and the absence of direct impact of additional contextual factors on time-to-recidivism or time-to-custody. We state it formally below.

Lemma 1. $\forall t > 0, m \in \{$*low, medium, high*$\}$, *if* $\mathbb{P}[\tau > t | D = \textit{majority}, M = m] \neq \mathbb{P}[\tau > t | D = \textit{minority}, M = m]$ *at significance level* α, *then we reject the null hypothesis* H_0 *that the context* U *does not directly affect time-to-recidivism* τ *at the* $1 - \alpha$ *confidence level.*

Lemma 1 lets us conclude whether the contextual factors directly affect time-to-recidivism when we see different no-recidivism curves for different races in the same algorithmic risk assessment decision groups. We explain below how to perform such an evaluation when the actual time-to-recidivism can be masked due to censoring.

C More Details on Empirical Test

C.1 Empirical Test

One potential challenge in directly using Theorem 1 and Lemma 1 is that we often cannot observe the true time-to-recidivism τ for all individuals, only a lower bound of τ. This occurs because some individuals return to custody for non-criminal violations, like missing probation meetings, before any potential reoffense - a phenomenon known as censoring in survival analysis. Traditional statistical tests that ignore censoring could produce biased results, as mistakenly using time-to-custody shall underestimate the true time-to-recidivism. Survival analysis methods are specifically designed to handle such censored data by properly accounting for both observed recidivism events and censored observations, thereby allowing us to decide if we have enough empirical evidence to reject the null hypothesis H_0 or not.

Under the null hypothesis, individuals of different races but the same algorithmic risk score group should have identical survival curves - that is, their probability of remaining arrest-free should be the same at all time points. To test this hypothesis while properly accounting for censoring, we employ the non-parametric log-rank test under assumptions made in Appendix C, which compares the entire survival curve rather than outcomes at a single time point.

Test Statistic. Let $O_{d,m}$ denote the observed rearrests for each race d in risk assessment group m across all event times and $E_{d,m}$ the expected rearrests similarly. The log-rank test statistic compares $O_{d,m}$ and $E_{d,m}$: $\chi^2 = \frac{(O_{\text{majority},m} - E_{\text{majority},m})^2}{Var(O_{\text{majority},m} - E_{\text{majority},m})}$ where $Var(O_{\text{majority},m} - E_{\text{majority},m}) = \sum_t \frac{N_{\text{majority},m,t} N_{\text{minority},m,t} O_{m,t}(N_{m,t} - O_{m,t})}{N^2_{m,t}(N_{m,t}-1)}$.

The test statistic χ^2 follows a chi-square distribution under the null hypothesis of one degree of freedom. Statistical significance is determined by calculating the corresponding p-value. If p-value < 0.05, we find enough evidence supporting the recidivism curves across racial groups are significantly different from each other, thus rejecting the null hypothesis that the risk scores are sufficient to explain the observed disparities and additional contextual factors do not directly affect recidivism; if p-value $\geq$ 0.05, we do not find sufficient evidence supporting the recidivism curves across racial groups are significantly different from each other, thus failing to reject the null hypothesis that additional contextual factors do not directly affect recidivism.

D More Details on Empirics

Having developed a causal framework and a format empirical test for analyzing racial disparities in recidivism, in this section, we use the COMPAS dataset collected by ProPublica to evaluate the extent to which the observed disparities can be explained by algorithmic risk scores alone and the role of additional contextual factors in our framework. At its core, we hope to evaluate to what extent do observed racial disparities in recidivism stem from algorithmic bias versus broader contextual factors? Our causal framework suggests that if disparities persist even after controlling for algorithmic risk scores, this would indicate the presence of additional unmeasured influences on recidivism trajectories. Specifically, we apply the empirical test developed in Section C.1 to examine whether and when racial disparities emerge in time-to-recidivism patterns. This allows us to assess not just the existence of contextual effects, but also their temporal dynamics - whether disparities appear immediately post-release or develop over longer follow-up periods. Such temporal patterns can provide insight into how structural inequalities may compound over time to shape recidivism outcomes.

D.1 Data Description

We preprocess the dataset to exclude cases with missing key variables, such as recidivism status or risk scores. Additionally, the COMPAS risk scores are categorized into three levels—*low* (1-4), *medium* (5-7), and *high* (8-10) —representing perceived recidivism risk, which serves as a proxy for algorithm-based criminal justice system decisions. We also distinguish between two key outcomes: rearrest for criminal offenses (the primary recidivism event) and return to custody for non-criminal violations (treated as censoring events in our analysis).

It is important to note that the COMPAS dataset, while widely used, has limitations inherent to criminal justice data. These include potential sampling biases, variations in law enforcement practices, and the absence of certain contextual factors such as socioeconomic status or access to community support. Additionally, the COMPAS dataset reflects only a specific jurisdiction—Broward County, Florida—which may limit generalizability to other regions with differing criminal justice practices.

D.2 Results

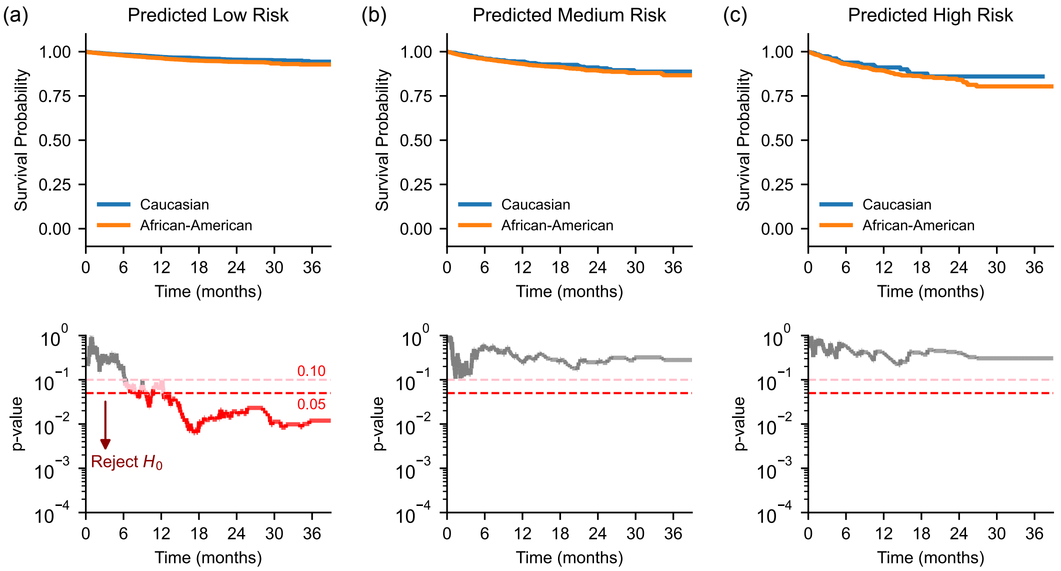

Figure 3: Survival analysis of recidivism patterns across racial groups and COMPAS **violent** recidivism risk groups. The subplots display survival curves and statistical significance analysis: (a) survival curves for Caucasian defendants, (b) survival curves for African-American defendants, and (c) corresponding p-values from log-rank tests over time. Gray ($p > 0.1$) indicates insufficient evidence of racial differences, light pink ($0.05 < p \leq 0.1$) indicates marginal differences, and red ($p \leq 0.05$) indicates significant differences.

We repeat the same empirical analysis for specific COMPAS recidivism risk scores and violent recidivism risk scores, i.e. scores 0 through 9 rather than quantized to $\{low, medium, high\}$. The results are shown in Figure 5 and 6 respectively.

Our analysis reveals distinct temporal patterns that vary with assigned risk scores. For individuals receiving all risk score except 3 or 4, the no-recidivism curves for Caucasian and African-American defendants remain similar throughout the follow-up period. Log-rank tests confirm this observation, showing no statistically significant differences between racial groups (p > 0.1). This result might also be due to limited data in each risk score.

However, a markedly different pattern emerges among individuals who received recidivism risk score 3 or 4 (in the low risk score group). While recidivism trajectories are initially similar between racial groups, significant disparities begin to appear after approximately seven months of follow-up (p < 0.05). Beyond this point, African-American defendants show a faster decline in their no-recidivism probability compared to Caucasian defendants who were assessed with the same low risk scores.

To examine potential racial disparities in recidivism patterns, we conducted survival analyses stratified by COMPAS risk groups. There are two major types of risk scores predicted by the COMPAS algorithm: risk for recidivism and risk for violent recidivism. Figure 2 and 3 present two complementary visualizations for each risk group and type: no-recidivism curves showing the proportion of individuals who have not recidivated over time, and corresponding statistical significance levels from log-rank tests comparing racial groups. Gray-shaded p-values indicate insufficient evidence to distinguish time-to-recidivism patterns between groups; light pink signifies marginal differences (significance level of 0.1), while red indicates significant differences (p-value < 0.05).

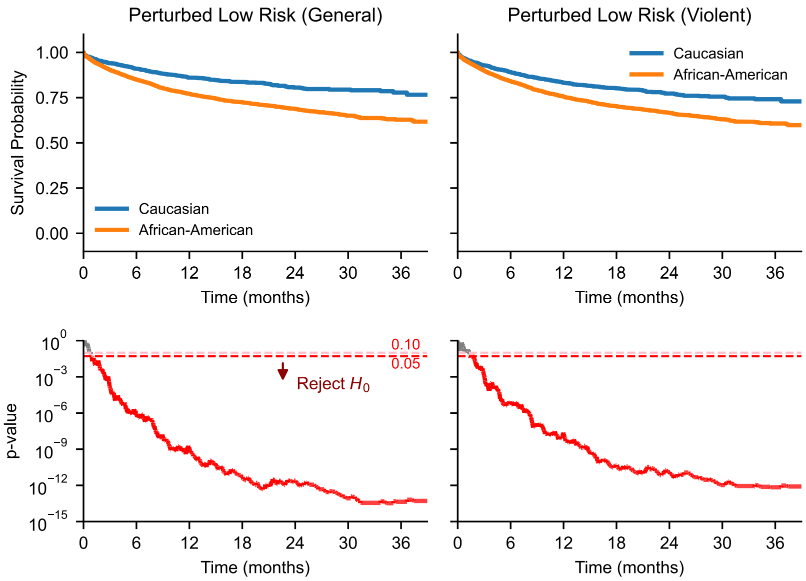

Figure 4: Survival analysis of recidivism patterns across racial groups and **perturbed** Low COMPAS recidivism risk groups. 10% of the African-Americans' risk scores are randomly shifted lower and 10% of the Caucasians' risk scroes are randomly shifted higher. The subplots display survival curves and statistical significance analysis: (a) survival curves for Caucasian defendants, (b) survival curves for African-American defendants, and (c) corresponding p-values from log-rank tests over time. Gray ($p > 0.1$) indicates insufficient evidence of racial differences, light pink ($0.05 < p \leq 0.1$) indicates marginal differences, and red ($p \leq 0.05$) indicates significant differences.

Our analysis reveals distinct temporal patterns across risk categories. For individuals classified as medium or high-risk by either risk of recidivism or risk of violent recidivism, the no-recidivism curves for Caucasian and African-American defendants remain similar throughout the follow-up period. Log-rank tests confirm this observation, showing no statistically significant differences between racial groups ($p > 0.1$). This suggests that within these higher risk categories, the algorithmic risk scores effectively capture recidivism patterns across racial groups.

However, a markedly different pattern emerges among individuals classified as low-risk by either risk of recidivism or risk of violent recidivism. While recidivism trajectories are similar between racial groups within a short follow-up period, significant disparities begin to appear with longer periods approximately seven months of follow-up ($p < 0.05$). Beyond this point, African-American defendants show a faster decline in their no-recidivism probability compared to Caucasian defendants who received identical risk scores.

The log-rank test results provide formal statistical evidence for these observations. For medium and high-risk groups, we fail to reject the null hypothesis that contextual factors has no direct effect on recidivism timing. However, for the low-risk group, we reject this null hypothesis after the seven-month mark, indicating that factors beyond the algorithmic risk assessment significantly influence recidivism patterns.

D.3 Discussion: Socioeconomic Contextual Influences on Recidivism

While initial short-term analyses suggest comparable recidivism outcomes across races, disparities become more pronounced over extended follow-up periods, which indicates the growing influence of non-algorithmic factors that the algorithm does not - and perhaps cannot - account for. The fact that disparities emerge most strongly in the low-risk group is especially concerning, as these individuals might otherwise have the highest potential for successful reintegration.

We argue that one highly plausible source of these non-algorithmic influences is socioeconomic disadvantage, including barriers to long-term housing, food security, and stable employment. This

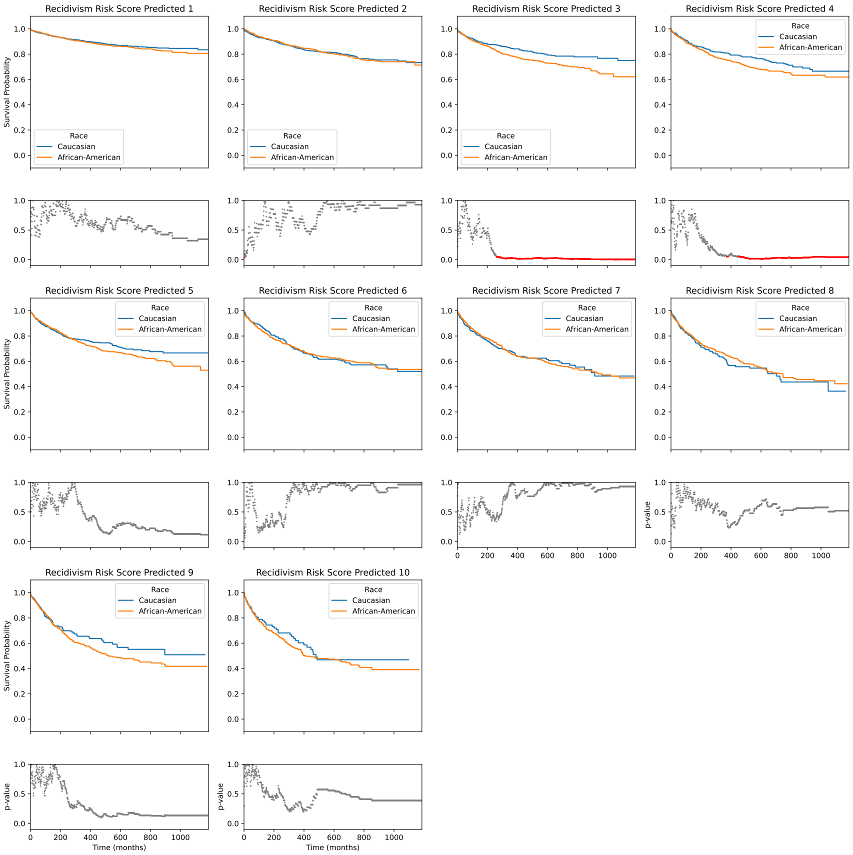

Figure 5: Survival analysis of recidivism patterns across racial scores and COMPAS recidivism risk groups. The subplots display survival curves and statistical significance analysis: (a) survival curves for Caucasian defendants, (b) survival curves for African-American defendants, and (c) corresponding p-values from log-rank tests over time. Gray ($p > 0.1$) indicates insufficient evidence of racial differences, light pink ($0.05 < p \leq 0.1$) indicates marginal differences, and red ($p \leq 0.05$) indicates significant differences.

interpretation aligns with findings from [4], who emphasize the critical role of targeted support services in mitigating recidivism risks among disadvantaged groups. The differential impact of societal contexts on minority individuals, particularly concerning access to essential services like housing and employment, reinforces the necessity of contextualizing algorithmic predictions within broader socioeconomic frameworks.

Within the context of racial disparities, it becomes apparent that counterfactual fairness, as defined earlier in this paper, may hold in the short term but falters over longer periods due to cumulative and compounding societal inequalities. The empirical evidence highlights the complex interplay between risk assessment tools and broader structural factors, challenging policymakers to implement comprehensive reforms that extend beyond algorithmic fairness.

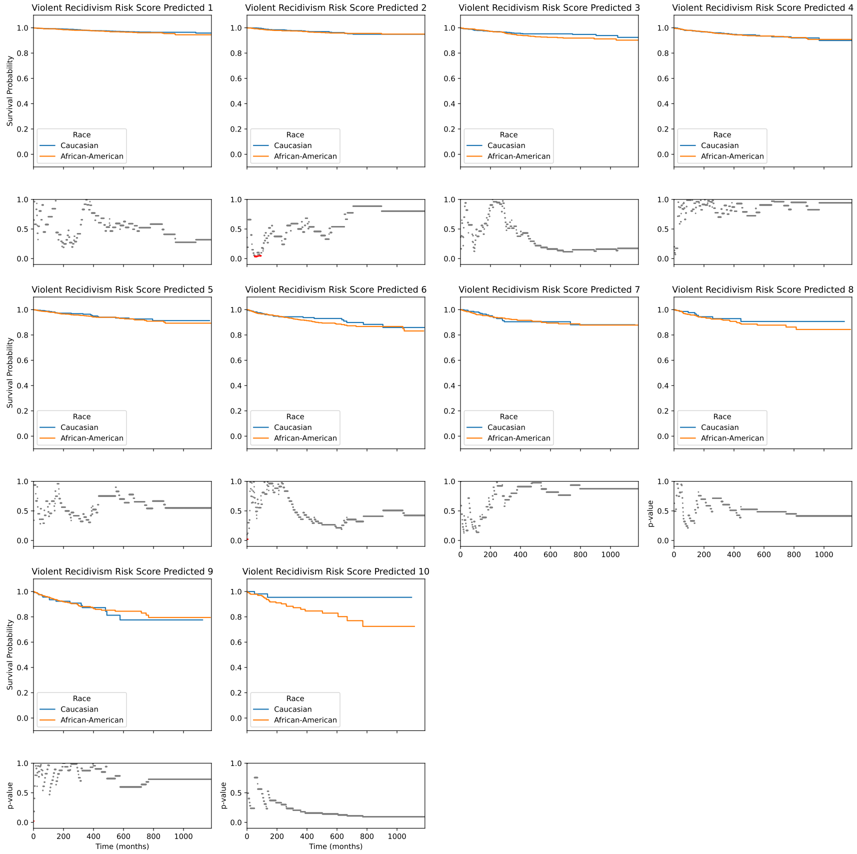

Figure 6: Survival analysis of recidivism patterns across racial scores and COMPAS **violent** recidivism risk scoress. The subplots display survival curves and statistical significance analysis: (a) survival curves for Caucasian defendants, (b) survival curves for African-American defendants, and (c) corresponding p-values from log-rank tests over time. Gray ($p > 0.1$) indicates insufficient evidence of racial differences, light pink ($0.05 < p \leq 0.1$) indicates marginal differences, and red ($p \leq 0.05$) indicates significant differences.