Safe to Hire: Predicting Recidivism Risk for Job Candidates with Criminal Records

Elizabeth C. Chase*†, Shawn D. Bushway †‡,
Bethany Saunders-Medina†, and Lane F. Burgette†

January 17, 2025

Abstract

In the United States, the use of criminal history information in employment decision-making is ubiquitous. However, employers' decision-making about job candidates' criminal records is often nontransparent and inconsistent, with disproportionate effects on Black and Hispanic Americans. Here, we consider whether statistical models can improve decision-making about job candidates with criminal records. We review existing approaches and policy guidance on the use of criminal records in employment decision-making. Then, using data from seven states from the Criminal Justice Administrative Records System (CJARS), spanning 1992-2021, we build a Cox proportional hazards model to predict the risk of recidivism based on a job candidate's criminal record. We assess the predictive performance, fairness, and trans-state portability of this model. We find that our candidate model outperforms some existing approaches, although challenges remain in the domains of fairness and usability in practice.

Keywords: criminal justice, criminal background checks, survival analysis, algorithmic fairness, redemption

*The authors gratefully acknowledge Arnold Ventures, for funding this project, and the staff of the Criminal Justice Administrative Records System (CJARS), particularly Jordan Papp, for assistance with the CJARS data. We also thank Dulani Woods at RAND for assistance with data preparation and Greg Baumann and Lou Mariano at RAND for helpful feedback and suggestions.

†The RAND Corporation

‡Department of Public Administration and Policy, University at Albany, SUNY

1 Introduction

Almost a third of Americans have a criminal record [Bushway et al., 2022, Kuhn, 2019]. For these Americans, the consequences of their offense often extend beyond any court-imposed penalty, and are especially acute in the job market. Employers are averse to hiring people with criminal records [Baier, 2020, Stoll and Bushway, 2008, Bushway et al., 2022, Pager, 2003, Agan and Starr, 2017], with between 16-60% of employers reporting that they would not hire anyone with a criminal record [Cavico et al., 2014, Levashina et al., 2017, Bushway and Pickett, 2024, Cullen et al., 2023]. For this and other reasons, Americans with criminal records are more likely to be unemployed than people without records [Couloute and Kopf, 2018].

The consequences of reduced hiring of people with criminal records extend to society as a whole. Research suggests that employment can reduce the risk of further crime, particularly for young men [Cavico et al., 2014, Denver et al., 2017]. In a tight labor market, refusing to consider the 30% of the U.S. population with criminal records further reduces an already-limited candidate pool [Ferguson, 2024]. By not seriously considering people with criminal records for good jobs, society leaves a useful labor market largely untapped, while potentially increasing these individuals' risk of recidivism (i.e., reoffending).

Here, we consider how criminal background information is currently used in U.S. hiring, and we propose a quantitative framework to assist employers with assessing the recidivism risk of individuals with criminal records. We focus on four main research questions:

1. **How is criminal history information currently used in employment decision-making?** What regulations govern these decisions, what are some examples of risk assessments currently in use, and what evidence supports these assessments?

2. **Can we develop a better recidivism risk assessment for employers?** By

"better," we mean more accurate, more usable, more transparent, more fair, and compliant with U.S. federal guidance on the use of criminal history information for employment decision-making.

3. **What are the properties of our recidivism risk assessment?** How does it make predictions, and what kinds of predictions does it generate? How well does it predict future crime? Is it fair? How does it compare to existing approaches?

4. **How might we further improve our recidivism risk assessment?** What are the missing components of this endeavor, in terms of data, methodology, and policy?

Overall, we have several key takeaways. First, we find that the majority of individuals with criminal records *will not reoffend.* This is true even over long periods of follow-up. We will discuss why other research has found higher rates of recidivism than what we report here, but we find that most people with criminal records will desist from crime.

Second, it is possible to develop a substantially better recidivism risk assessment for hiring decisions than the methods that employers typically use, and we have done so here. Especially when employers use ad-hoc approaches to assess risk, there is greater room for individual biases and prejudices to drive hiring decisions. Moreover, because we find that recidivism risk drops more quickly over time and to lower levels than conventional wisdom would suggest, employers may be more conservative in their recidivism hiring choices than is warranted. We believe our risk assessment framework is an improvement in terms of accuracy, transparency, fairness, and legal compliance.

To develop this risk assessment, we make use of the Criminal Justice Administrative Records System (CJARS), the most comprehensive criminal history data available in the United States [Finlay and Mueller-Smith, 2022, Finlay et al., 2022], which enables us to consider conviction data (rather than arrests) and, critically, information on incarceration,

which much previous research has been unable to do [Bushway and Kalra, 2021, Blumstein and Nakamura, 2009, Bushway et al., 2022]. Our model is reasonably transparent and provides uncertainty estimates. And our data enable us to consider model performance with regards to statistical accuracy, algorithmic fairness, and model portability across states.

But, third: although our recidivism risk assessment improves upon current practice, more work remains to produce a risk assessment that truly meets the needs and rights of employers, people with criminal records, and the American public. This future work primarily falls in the domains of data collection and policy. With regards to data, although CJARS has undertaken the monumental endeavor of cleaning and harmonizing American criminal justice data, more work is needed to acquire, check, and clean criminal records data for recidivism risk assessment. With regards to policy, work remains to clarify how a quantitative recidivism risk assessment might be used within employment decision-making, what level of recidivism risk employers should be willing (or legally required) to tolerate of their employees, and continued work to ensure that employers have access to the criminal records information they are legally allowed and encouraged to consider as part of hiring—while restricting their access to data they are not allowed to use.

The remainder of the paper is as follows. We will address our first research question in the next section, which reviews existing policy and approaches for recidivism risk assessment. Then, we will present our approach to quantifying recidivism risk. We will investigate our model's performance and properties, answering research questions two and three. Finally, we will close with a discussion that considers research question four: the limitations of our work and the missing pieces needed to enable better hiring decisions for Americans with criminal records.

2 Existing Approaches for Employment Recidivism Risk Assessment

Even as the percentage of Americans with criminal records is the highest in our history, and the stigma surrounding criminal records severe, it is easier than ever before to access criminal history information. The United States has an open criminal records system [Bushway and Kalra, 2021], making it straightforward to access a person's criminal history, a process that has accelerated in the digital age [Blumstein and Nakamura, 2009]. Numerous private companies perform criminal background checks (CBCs) for employers [Stoll and Bushway, 2008, Smith, 2014], with roughly 90% of employers now using some kind of CBC prior to hiring [Cavico et al., 2014, Kurlychek et al., 2019, Bushway and Kalra, 2021]. At the same time, legal requirements for CBCs for particular jobs have rapidly expanded in the last 40 years [McElhattan, 2022].

In recognition of the enormous impact that having a criminal record can have on employment outcomes—and the fact that such impacts are particularly acute for Black and Hispanic Americans, who are much more likely to have a criminal record than White Americans, for a variety of reasons [Cavico et al., 2014]—the Equal Employment Opportunity Commission (EEOC) has taken an active stance on regulating the use of criminal record information in hiring decisions. In their most recent guidance, in 2012, the EEOC listed factors that employers should consider in their decision, such as the crime type and gravity, the harm caused by the crime, the number of offenses, the length of time since the last offense, the age at offense, the nature of the job, the relevance of the crime to the job sought, and rehabilitation efforts [Equal Employment Opportunity Commission, 2012]. Critically, they advised employers to only consider convictions—not arrests—because arrests are not

indicative of guilt and disproportionately burden Black and Hispanic Americans [Cavico et al., 2014, Neil and Zanger-Tishler, 2025]. They required employers to develop a "narrowly tailored" policy for excluding particular types of criminal records that had to involve an "individualized assessment." In further guidance, the EEOC recommended that employers first undertake a "targeted screen" of job candidates' criminal records, followed by an individualized assessment [Equal Employment Opportunity Commission, 2013]. The decision-making process must be transparent, and applicants who are excluded because of their criminal history are entitled to know why. As an overarching principle, if the use of criminal history in employment decision-making has a disparate impact on a protected category of Americans (which, because of the racialized nature of criminal justice in America, it almost certainly will), then the use of the criminal history information must be job related and consistent with business necessity, and there must not exist an equally effective but less discriminatory alternative [Equal Employment Opportunity Commission, 2012, Jacqueline Ramos, et al., v. Walmart, Inc., 2021]. Informally, the EEOC defines disparate impact by the "four-fifths rule": if hiring rates for one protected group (e.g. Black applicants) under a particular selection criterion are less than four-fifths the hiring rates for another group (e.g. White applicants), then the selection criterion has disparate impact [Equal Employment Opportunity Commission, 2023].

In response to this guidance and federal and state legislation, many large employers have converged to the standardized hiring process shown in Figure 1. Employers assess job applicants without knowledge of the applicant's criminal background. After choosing to issue a conditional job offer, employers then request a CBC, which undergoes a targeted screening process. Applicants who pass the targeted screen receive a finalized job offer; applicants who do not pass the targeted screen receive a letter informing them of the situation

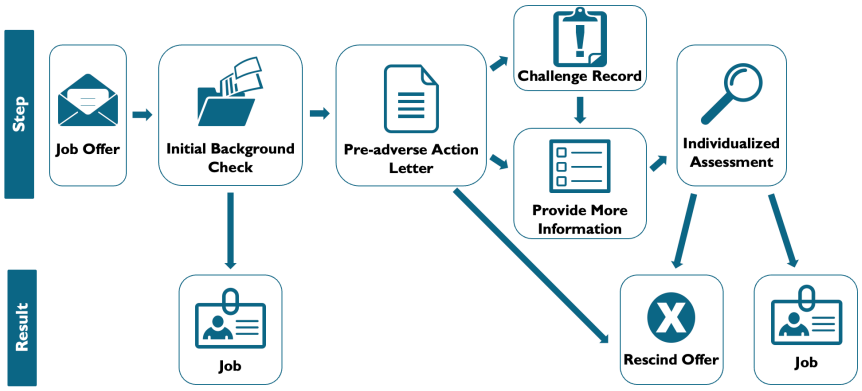


Figure 1: Schematic representation of the standardized criminal background check process used for hiring by many U.S. employers. Figure taken from Corporate Coalition of Chicago, "Hiring Landscape for Justice-Impacted Individuals."

and requesting more information about their case, which is used to guide the individualized assessment. In some cases, candidates who do not respond to this letter may immediately have their offer rescinded; others may progress to the individualized assessment despite not providing additional information. After the individualized assessment, candidates either receive a finalized job offer or the conditional job offer is rescinded.

Within this framework, we see a quantitative recidivism risk assessment as fitting most naturally in the targeted screen step. Individuals who are flagged as higher risk by the targeted screen can then undergo a non-algorithmic individualized assessment. The EEOC's guidance provides some constraints and guiding principles for building this targeted screening model. The factors highlighted by the EEOC (age at conviction, number of convictions, etc.) can be seen as predictors in the model. That model should not consider information

on arrests, only on convictions and incarcerations. It also must not consider information on any protected category, such as race, gender, or nationality. The differential impact of model predictions for different Americans (Black or White, men or women, etc.) should be assessed and quantified, and if another model exists that has equivalently good predictive performance but less differential impact, then that other model should be preferred. Perhaps such a model should only predict particular types of offenses that are relevant to a given job. The model needs to be interpretable, so that job applicants can check the results of the model prior to the individualized assessment and understand which elements of their criminal record are being weighted most heavily.

That, at least, is the EEOC guidance. But in practice, how do employers carry out the targeted screen? A standard CBC will contain information on lifetime conviction history and any arrests in the past 7 years [Cavico et al., 2014]. It is not always clear how arrests are connected to convictions, making it difficult to tell if an arrest resulted in a conviction or a ruling of not guilty. Although employers are not supposed to consider arrests that did not result in a conviction, many do [Kuhn, 2019, Bushway et al., 2022, Levashina et al., 2017]. There is no information on a CBC about incarceration or rehabilitation, making it difficult for employers to consider these factors unless they request them of applicants. Because of heterogeneity in plea bargains, sentencing, and crime classification, people guilty of identical criminal behavior may have different records [Kuhn, 2019]. In a survey of human resource professionals—many of whom had had training on assessing CBCs—many still reported that they were frequently confused by CBC terminology [Kuhn, 2019].

Those are the data challenges; in statistical terms, the issues caused by the inputs. Even if we imagined that employers had perfect CBC data, how well would they do at assessing risk and making decisions? In statistical terms, how's the model? Human resource profes-

sionals report a lack of systematic approaches to decision-making about CBCs, usually favoring black box approaches (e.g. employer discretion) unless legal mandates lay out clearly defined rules [Kuhn, 2019, Bushway and Kalra, 2021]. A cottage industry of companies has sprung up to assist employers with developing their targeted screenings [APTMetrics, 2024, VeriFirst Background Screening, LLC, 2024, Cisive, 2021, Yardstik, 2024], promising to "help you identify appropriate job-related criteria for excluding criminal offenders from hire" [APTMetrics, 2024] and offering templates for targeted screening rubrics that can be tailored by the employer [VeriFirst Background Screening, LLC, 2024]. It is not clear what percentage of employers use such services or have independently developed their own formalized targeted screening criteria, but there is evidence that a nontrivial percentage of employers have no formalized targeted screening, instead making ad-hoc decisions on a case-by-case basis [Cisive, 2021, Kuhn, 2019, Levashina et al., 2017]. (We refer to this as the "lawyers in a room" targeted screening approach).

Research suggests that there is major heterogeneity in opinions about which criminal records should be disqualifying, even for the same job, meaning that screening results may depend a great deal on which employer (or which lawyer in a room) is carrying out the screening. For example, in a survey in which reviewers were given a criminal history to review for a particular job, 20% of reviewers said they would reject the applicant on the basis of their criminal history, while an additional 30% said that the criminal history posed no problem "at all" [Kuhn, 2019]. Despite EEOC guidance that employers should consider crime type and severity, most employers do not appear to do so, when presented with criminal histories to review [Kuhn, 2019]. They do seem to consider time since last offense, with most employers gravitating towards a 10 year threshold for sufficient redemption [Kuhn, 2019].

Most of the above findings are backed out from hiring surveys, rather than reviewing the targeted screening approaches directly: targeted screening approaches are rarely made public, because it exposes the employer to legal risk if it is demonstrated that an equally effective but less discriminatory alternative exists. Nontransparent, black box approaches are the norm. (Kurlychek et al. [2019] is a notable exception.) Given this, it is challenging to identify competitors to which we might compare our model, because most decision-making approaches are confidential. Instead, we compare our model to some of the criminal history requirements imposed by law for particular jobs. These requirements bar individuals with particular types of convictions (usually either felonies or violent offenses) within a certain time window (often 7 or 10 years, or a lifetime ban) [Blumstein and Nakamura, 2009, Bushway et al., 2022, Kuhn, 2019]. We will compare to approaches of this type.

3 Material and Methods

3.1 Data

Our data come from the Criminal Justice Administrative Records System (CJARS), a project run by the University of Michigan in partnership with the U.S. Census Bureau. CJARS' goal is to obtain, clean, and harmonize U.S. administrative criminal justice records [Finlay and Mueller-Smith, 2022, Finlay et al., 2022]. Because the criminal justice system (and its willingness to share its data) varies from state to state, if not county to county, data quality and availability is not consistent across the United States. At present, CJARS contains over 3 billion records on 44 million unique individuals across 34 states.

Note that recidivism is a time-to-event outcome, i.e., time to next offense. Therefore, it is critical to construct a comprehensive criminal history timeline of when a given individual

experienced convictions and incarcerations. For our analysis, we needed information on two types of criminal justice interactions: convictions and incarcerations, ideally with long follow-up. We focus on convictions because convictions, unlike arrests, are more indicative of engaging in criminal behavior [Neil and Zanger-Tishler, 2025]. This is one reason that the results we present here may differ from previous research that has used arrest as an endpoint; recidivism rates are lower when the focus is convictions rather than arrests. We need incarceration information because an individual cannot apply for a job while incarcerated and thus is not part of our target population; these individuals also are not truly at-risk of reoffending in the typical sense, because they are in prison. Not adjusting for incarceration time (and thus treating incarcerated individuals as if they were free in society) risks biasing the estimated recidivism rate downwards, because these individuals might have reoffended if they were not in prison. With the exception of Bushway et al. [2022], no previous studies to our knowledge have been able to adjust for incarceration when modeling recidivism risk, which is another reason the results we present here may differ from previous work. Some studies have attempted to address the challenge of incarceration by only focusing on prison-exit cohorts, but not all convictions result in incarceration, and thus prison-exit cohorts are a nonrepresentative sample. Therefore, our results will also differ from prison exit-cohort studies.

With those requirements in place (comprehensive conviction and incarceration data on the whole state, with long follow-up times) we were able to focus on seven states in CJARS: Arizona, Florida, Minnesota, Oklahoma, North Carolina, Texas, Wisconsin. Years of follow-up on each state are given in Table 1.

Table 1: Years of follow-up for each state included in this analysis.

State	Years
Arizona	1997-2020
Florida	2005-2021
Minnesota	2009-2021
Oklahoma	2000-2017
North Carolina	2011-2020
Texas	1992-2012
Wisconsin	2000-2019

3.2 Data Processing

Because data construction and quality has a large impact in this setting, we will describe our data cleaning process, although full details are deferred to Appendix 1. For each of the seven states, we identified all individuals in the CJARS data who had a conviction in that state. We then pulled all CJARS records pertaining to this individual (even if the records were from states outside of our set of seven) to construct a maximally complete criminal history for them. This gives rise to the first data-based limitation of our analysis: because CJARS does not have comprehensive data on all 50 states, we are missing records for individuals who migrated out of CJARS coverage. The rate of out-of-state migration varies depending on the state, but it is nontrivial [Durose et al., 2015]. We expect out-of-state migration to bias our recidivism estimates to be lower than they otherwise should be, because we are missing offenses.

CJARS has linked all records from the same individual. However, they have not linked each incarceration sentence to a given offense; that work is still under development [per-

sonal communication from Jordan Papp, Project Manager at CJARS, to authors, Dec. 1, 2023]. As a result, the information we obtain for a given individual is one dataset with all of the convictions in CJARS for that individual, and a second dataset with all of the incarcerations in CJARS for that individual. There is no mapping across the two datasets, and multiple convictions that were handed down for the same criminal episode are not linked, either. This results in a second challenge of our work: identifying convictions that correspond to the same criminal episode, and linking convictions to incarceration. For grouping convictions, we used a non-rolling threshold of 30 days: all convictions that fell within a 30 day window were considered to be part of a single episode. For matching incarcerations to convictions, we mapped the incarceration to the most recent conviction, and assumed that any intervening time between conviction and reported incarceration start date was spent in prison. Because CJARS does not include county jail data, we think this is plausible.

We use this information to construct a unified dataset in which individuals have multiple rows, with each row corresponding to a single crime episode. That row will have information on how many convictions and which crime types resulted from the crime episode, and whether any incarceration time was imposed, along with information on the number of prior convictions, prior crime types, and prior incarceration time. Each row also contains the length of time spent free-in-society prior to the next offense, with a censoring indicator for whether another recidivism event was observed during follow-up or not. We restrict to convictions between the ages of 15-60, because these are prime employment years.

Finally, we need to select a single crime episode per individual. Including all of the conviction episodes upweights individuals with extensive criminal histories—those with very high recidivism rates [Neil and Zanger-Tishler, 2025, Bushway et al., 2022]. To obtain an accurate reflection of recidivism risk among all individuals with convictions, we should

only include a single record per person. Previous work has randomly sampled a single record proportional to number of records [Bushway et al., 2022]; here, we take a different approach, instead sampling a single record proportional to *time spent free in society*. For each individual, we calculate their total time spent free in society from their first offense to the end of follow-up. We then weight each offense by what proportion of their total time spent free in society was spent after that offense. In the data we observe that individuals sometimes go through high crime periods where they may have many criminal episodes in the space of 2-3 years, usually while young. If we sample proportional to number of records, we are more likely to sample a conviction during a high crime period, making the estimated recidivism risk appear higher than it may actually be, considering the perspective of a potential employer who would be unlikely to encounter an applicant in the short period of time between multiple rapid convictions. By sampling proportional to time spent free in society, we address this issue. To our knowledge we are the first to notice this problem, which is another reason our estimated recidivism rates differ from previous work. We direct interested readers to Supplemental Figure S2, which shows how the choice of sampling strategy affects estimated recidivism rates in the CJARS data.

After sampling a single criminal episode per individual, we have our final analytic sample: a single crime episode (row) per individual, with their criminal history up to that point included as covariates, and with a length of time to next offense and censoring indicator for whether the next offense was observed prior to end of follow-up. We then split the sample into training and validation samples.

3.3 Model Building

As candidate models, our initial plan was to fit a Cox proportional hazards (PH) model [Cox, 1972] and a Bayesian Additive Regression Tree (BART) [Chipman et al., 2010, Sparapani et al., 2016]. We regarded the Cox model as the interpretable and transparent option, while we regarded the BART as a benchmark near the upper bound on predictive performance. We hypothesized that the Cox model could provide comparable predictive performance to the BART, based on work by Liu et al. [2022] and previous research suggesting that the PH assumption may hold in this setting [Bushway et al., 2022]. We initially fit a Cox PH and a BART. Both models used the following set of predictors:

- Predictors on age: age at offense or release from prison, age at first conviction
- Information on most recent offense: whether it resulted in any felony convictions, whether it resulted in any misdemeanor convictions, crime type, how much incarceration time (if any) was sentenced
- Information on prior criminal record (before most recent offense): number of prior felonies, number of prior misdemeanors, number of prior convictions in each crime type category, number of prior incarcerations, total prior incarceration time

Additionally, the Cox PH model used a quadratic term for age and had interactions between age and whether the most recent offense resulted in a felony conviction, age and the number of prior felonies, age and the number of prior misdemeanors, and age and the amount of prior incarceration time. Models were fit separately for each state.

Unfortunately, given the size of our data, we were unable to get the BART to converge even in a subsample of only $n = 30,000$ records of our training data. This is in keeping with previous work on BART's convergence in large samples [Sparapani et al., 2021]. Therefore,

we switched to using a generalized boosted model (GBM) [Ridgeway, 2024], viewing this as a frequentist analogue to BART. Based on comparisons of the point estimates from the non-converged BART to the point estimates of GBM, we think this was a reasonable substitution. After deciding to switch to GBM, we fit our final GBM and Cox PH models in the training sample, separately for each state.

3.4 Evaluating Predictive Performance

We then evaluated predictive performance of each model in the validation sample. Note that in this setting, the predictive quantity of interest is the risk of being convicted (reconviction) in the next year, conditional on predictors and how many years have passed since the last offense. Because the GBM and Cox PH models produce estimated survival curves for each individual, $\hat{S}_i(t|X_i) = \hat{Pr}(T_i > t|X_i)$, we calculated:

$$\hat{S}_i(\tau+1|X_i, T_i > \tau) = \hat{Pr}(T_i > \tau+1|X_i, T_i > \tau) = \frac{\hat{Pr}(T_i > \tau+1, T_i > \tau|X_i)}{\hat{Pr}(T_i > \tau|X_i)} = \frac{\hat{S}_i(\tau+1|X_i)}{\hat{S}_i(\tau|X_i)}$$

where $\tau = 1, ..., M_s - 1$ is a grid of single years ranging from 1 to a year prior to the last year of follow-up, $M_s - 1$, in state $s, s = AZ, ..., WI$. We then calculated the risk as $\hat{R}_i(\tau+1) = 1 - \hat{S}_i(\tau+1|X_i, T_i > \tau)$. Individuals with higher values of $\hat{R}_i(\tau+1)$ are at greater risk of reconviction within the next year. Additionally, we generated predictions from our comparison methods: binary 0/1 predictions based on whether the individual had a felony conviction, a violent offense, or either a felony or violent conviction in the last 10 years. If the individual had a problematic charge within the last 10 years, their value of $\hat{R}_i(\tau+1) = 1$; if not, $\hat{R}_i(\tau+1) = 0$.

We used the time-dependent area under the curve (AUC) [Blanche et al., 2013] to evaluate model discrimination. For each value of $\tau = 1, ..., M_s - 1$, we identified individuals

in the validation sample with an observed survival time $> \tau$. We then calculated an AUC for their predicted risk of reconviction within the next year (by $\tau + 1$) compared to whether they were actually convicted by $\tau + 1$. We also examined calibration. For calibration, we again restricted the validation sample to individuals with an observed survival time $> \tau$. We calculated a predicted risk of reconviction by $\tau + 1$. We split the sample into deciles of predicted risk and calculated the mean predicted risk in each decile. We then compared the mean predicted risk to the observed risk within each decile using a calibration plot.

3.5 Evaluating Algorithmic Fairness

We considered fairness across race/ethnicity and sex. For brevity, we will refer to "race/ethnicity" as "race" in the remainder of the manuscript. We consider 3 primary race groups: non-Hispanic White (sometimes abbreviated to "White"), non-Hispanic Black (sometimes abbreviated to "Black"), and Hispanic individuals. While CJARS also identifies non-Hispanic Asian or Pacific Islander, American Indian or Alaska Native, and "Other race/ethnicity" individuals, sample sizes were small in these groups and resulted in very noisy estimates. Therefore, our fairness analysis only considers Black, White, and Hispanic individuals. For analyses by sex, we have male and female as the two options; CJARS does not include information on gender. We will use male/men and female/women interchangeably throughout the text, although we note that we are considering sex, not gender, in this analysis. Both race and sex are likely determined by an administrator, rather than via self-report.

As our primary metrics of fairness, we calculated race- and sex-stratified AUC and calibration plots [Cabreros et al., 2023, Corbett-Davies et al., 2023]. In light of the EEOC's four-fifths rule, we calculated the independence limited disparate impact [Cabreros et al., 2023], in which we compared the average value of $\hat{R}(\tau)$ across each race/sex group, to

estimate the ratio of predictions between groups.

The remainder of our algorithmic fairness analysis was to interrogate any disparities that we observed. We were concerned about two sources of bias: feature bias and label bias. Feature bias occurs when predictors in our model have differential effects on the outcome by race or sex. For example, is the effect of having two previous felonies on the risk of reconviction different for Black men than for White men? These differences could be due to differences in measurement of the predictors or differences in the impact of the predictor on the outcome. Label bias occurs when the outcome means something different for members of different groups. In our setting, label bias and feature bias are closely linked, in part because many of the features (predictors such as prior felonies, prior misdemeanors, prior incarceration time) are just earlier episodes of the labels (outcomes) being detected. As with feature bias, we worry about label bias if we think that a Black man's criminal behavior is more likely to result in conviction than a White man's equivalent criminal behavior, or if we think that the rate of false convictions is higher for Black men than for White men [Grunwald, 2023]. Feature bias is usually fixable; label bias is not [Neil and Zanger-Tishler, 2025]. As discussed, in our setting we expect feature and label bias to go hand-in-hand because of the links between our predictors and our outcome.

To assess feature bias, we used an approach from Corbett-Davies et al. [2023] and fit a Cox PH model in which we regressed time to next offense conditional on the linear predictor of each candidate model, a main effect of race and sex, and interactions between race/sex and the linear predictor, to see if there were differential effects of the linear predictor by race/sex. To further interrogate those differences, we refit our candidate models including race/sex as predictors, and (in the Cox PH model), interactions between race and sex and all other predictors [Neil and Zanger-Tishler, 2025, Corbett-Davies et al., 2023]. We

examined how predictive performance, coefficient/influence estimates, and risk predictions changed once race/sex were included in the model, and whether particular predictors were highly modified by race/sex.

3.6 Evaluating Portability

Finally, we examined state portability. In the model-building, validation, and algorithmic fairness analyses, we conducted all analyses stratified by state. But a single model that works at the national level would be of interest to large employers. Having a separate model for each state is difficult to explain and justify. Is it feasible to fit one national model?

To answer these questions, we used the state-specific models from three of our seven states: the model that on average produced the lowest risk predictions (Minnesota), the highest (North Carolina), and the median (Arizona). We then applied these three models to generate predictions in the other states, and we looked at how performance shifted, using the metrics described above. We also fit an unstratified model in our training data that used equal samples of data from all seven states. We explored the predictive performance and algorithmic bias of this nonstratified model.

4 Results

4.1 Descriptive Statistics and Estimated Recidivism Rates

Information on sample exclusions and data cleaning are provided in Appendices 1 and 2. Table 2 gives baseline descriptive statistics on the sample, stratified by state. We note major differences in some of these characteristics across state. Individuals are younger at the time of offense in AZ, TX, and WI, and age at first conviction is younger in NC and

TX than in the other states. Women are a greater share of the data in MN and NC than in the other states; there are also large shifts in the race distribution in different states.

Figure 2 provides the estimated recidivism rates over time, stratified by state. Once again, we note differences in recidivism rates across state, with the seven states split into two groups of recidivism rates: Texas and North Carolina, which have much higher recidivism rates, and the other five states, which have reduced recidivism rates. However, even in Texas and North Carolina, which have higher recidivism rates, we still observe that a full two-thirds of individuals with a criminal record will not have been reconvicted in the 10 years after their conviction. In the lower risk states (Arizona, Florida, Minnesota, Oklahoma, Wisconsin), almost 80% of individuals have not been reconvicted 10 years after their conviction. The vast majority of people with a criminal record will not be reconvicted.

4.2 Final Models

Effect estimates from the Cox PH model, stratified by state, are given in Figure 3. Predictor influence values from the GBM, stratified by state, are shown in Figure 4. Results from the Cox model were fairly concordant across states. Being younger at the most recent offense or at first conviction increased the risk of recidivism. Having a felony count on the most recent conviction also increased the risk of recidivism, as did increasing numbers of prior convictions. For the GBM, the most influential predictors were age at first conviction and number of prior convictions. These findings (both the Cox model and GBM) are in agreement with previous work by Blumstein and Nakamura [2009] and Bushway et al. [2022] on how particular predictors affect recidivism risk.

Figure 5 shows the median predicted risk and interquartile range (IQR) in each state. The vast majority of individuals are at low risk of reconviction, with more than half of

Table 2: Descriptive statistics on the cleaned recidivism modeling sample built from Criminal Justice Administrative Records System data (CJARS), stratified by state.

Characteristic	AZ	FL	MN	NC	OK	TX	WI
Number of offenders	988,046	2,589,547	615,405	935,505	466,987	3,181,495	732,265
Median age at conviction	30.7	32.1	31.5	32.4	32.0	29.1	29.9
Median age at first conviction	29.2	28.1	30.6	25.5	29.2	26.3	27.1
Race N (%)							
Non-Hispanic White	411,665	1,095,093	384,755	459,946	224,372	1,419,154	486,015
	(44.0%)	(42.9%)	(63.7%)	(49.3%)	(48.4%)	(44.7%)	(67.3%)
Non-Hispanic Black	70,551	845,047	142,369	370,287	172,111	750,084	153,948
	(7.5%)	(33.2%)	(23.6%)	(39.7 %)	(37.2%)	(23.6%)	(21.3%)
Hispanic	407,078	585,067	53,581	81,037	55,670	974,288	63,020
	(43.6%)	(22.9%)	(8.9%)	(8.7%)	(12.0%)	(30.7%)	(8.7%)
Other Race	45,381	22,688	23,669	22,313	11,087	33,825	19,518
	(4.8%)	(0.9%)	(3.9%)	(2.4%)	(2.4%)	(1.1%)	(2.8%)
Female N (%)	243,123	643,363	173,639	253,021	106,678	710,883	178,407
	(24.6%)	(25.2%)	(29.5%)	(27.1%)	(23.4%)	(22.4%)	(25.1%)

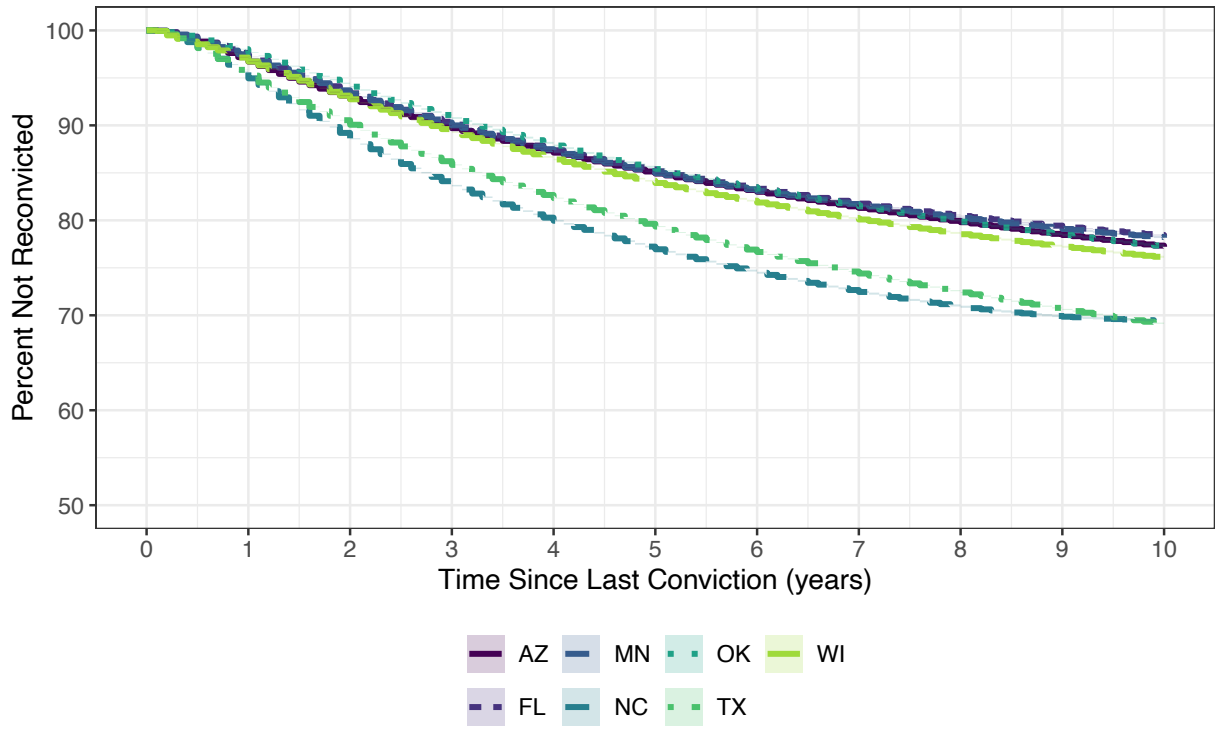


Figure 2: Estimated Kaplan-Meier curves for the risk of recidivism over time in a sample of $n = 9,509,250$ criminal records from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

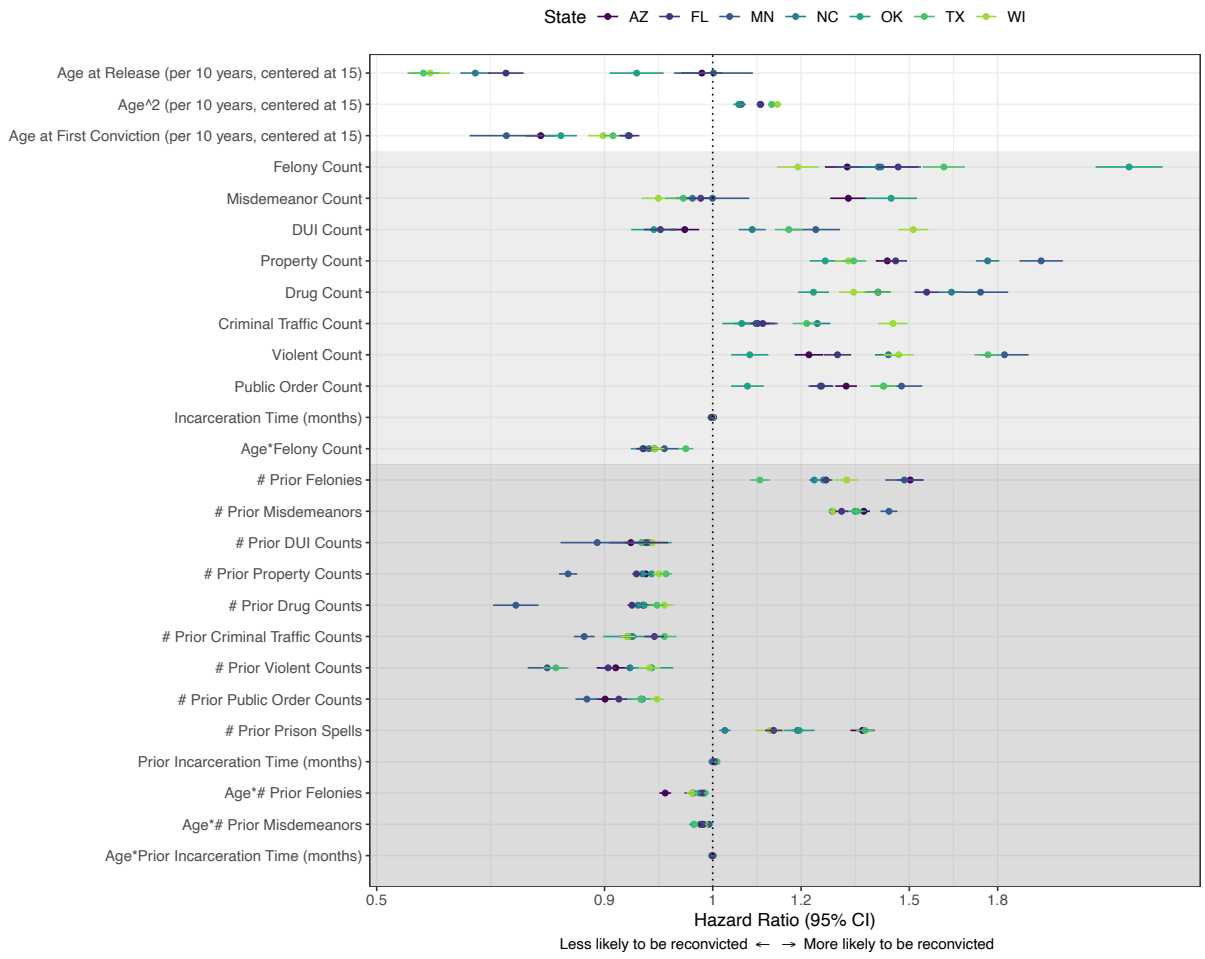


Figure 3: Hazard ratio and 95% confidence interval estimates for predicting recidivism using a Cox proportional hazards model fit in a training sample of criminal records from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

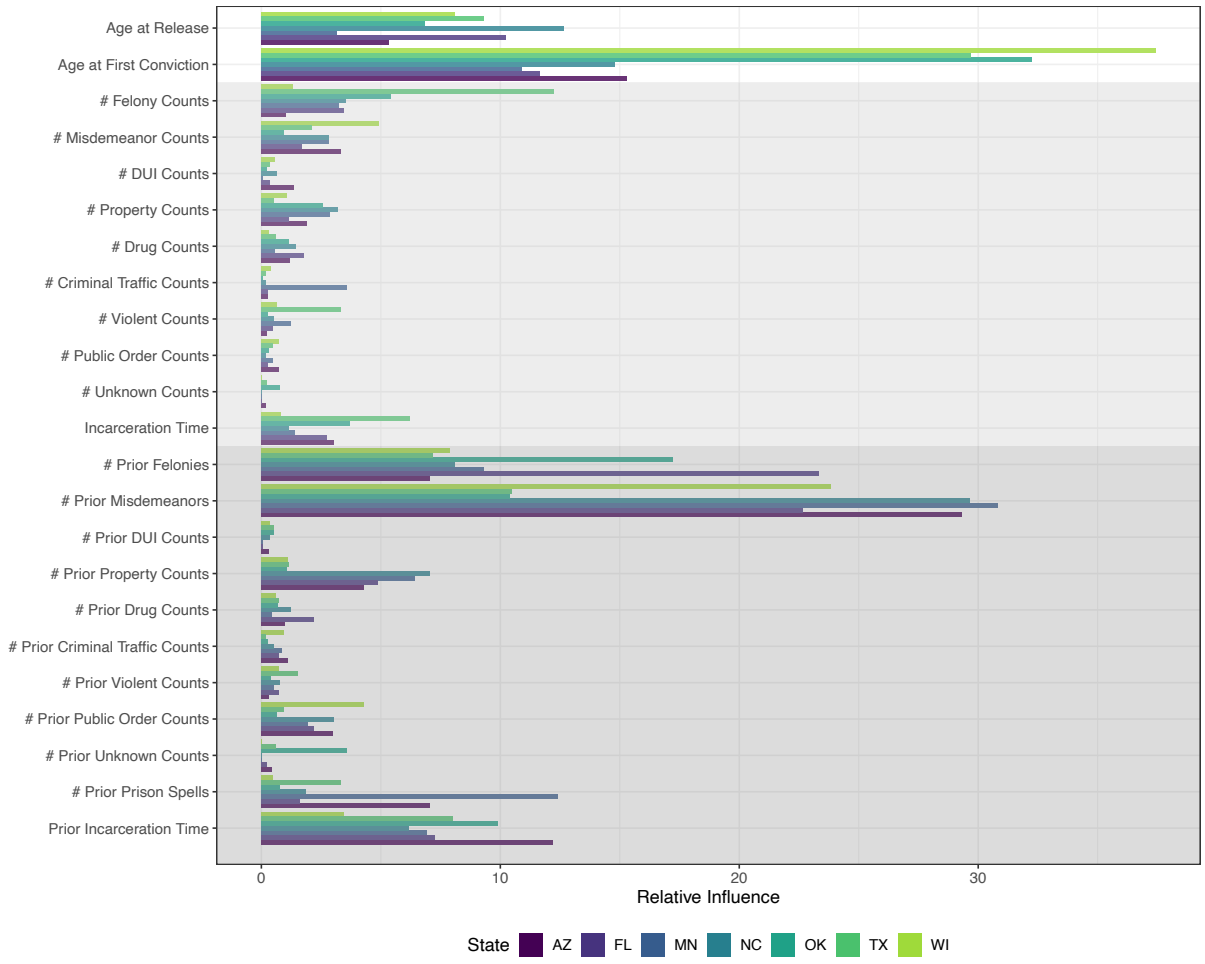


Figure 4: Relative influence of different characteristics for predicting recidivism using a generalized boosted model (GBM) fit in a training sample of criminal records from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

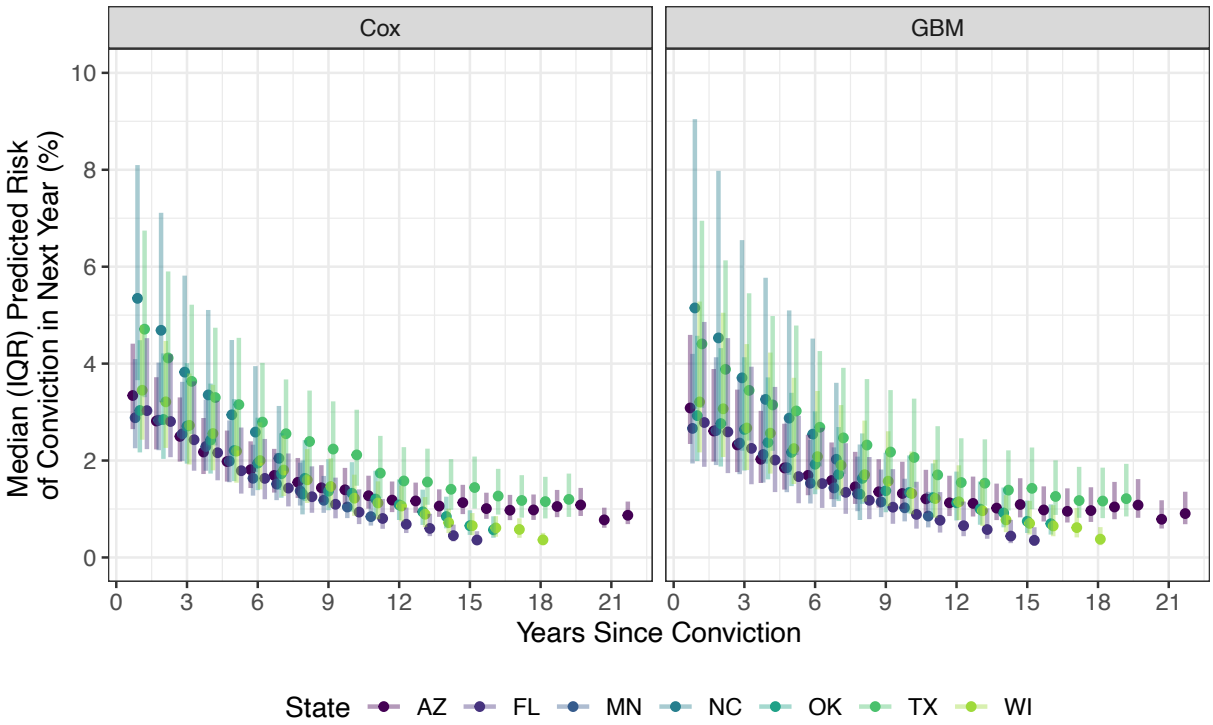


Figure 5: Median and interquartile range for the predicted recidivism risk as predicted by a Cox proportional hazards model (Cox) and generalized boosted model (GBM), stratified by state. Predictions generated on a validation sample of criminal records from the Criminal Justice Administrative Records System data (CJARS). Data range from 1992-2021, depending on the state.

individuals having $< 5\%$ chance of reconviction within the next year, even if it has been only a year since their most recent conviction. Risk of reconviction decreases as more time passes since the last conviction. We also note the different predictions across states: the lower predicted risks in Minnesota and the much higher predicted risk in North Carolina, with Arizona falling comfortably in the middle of the seven states.

We are also interested in the uncertainty around these predictions. Information on prediction interval width is summarized in Supplemental Figure S10. In general, most

predictions had 0.5-2 percentage points of uncertainty around them. Uncertainty was greater in Oklahoma and reduced in Arizona, Texas, and Florida—driven by the underlying sample sizes of the training data in each state.

4.3 Predictive Performance

Results on the AUC and calibration of each model are given in Appendix 3. Predictive performance was satisfactory for the Cox PH model and GBM. The GBM returned higher AUC values than the Cox PH model in all states, but these differences were small (e.g. 0.73 vs. 0.70). The GBM and Cox PH models outperformed the felony/violent offense decision rules at all timepoints. All models had reduced AUC as more time passed since last offense; after ten years have passed from the last conviction, no model is good, with the felony/violent offense decision rule exhibiting an AUC of 0.5 (as we would expect, given the ten year threshold), and the Cox/GBM models around 0.6. This leads to our first recommendation: while a recidivism risk assessment may be useful up to ten years after the last conviction, beyond that point it is of less utility and other approaches (such as a decision that all individuals are sufficiently low risk after ten years as to be safe to hire, or an individualized screen that considers rehabilitation, employment history, references, and current financial/family situation) should be preferred. Calibration (Figures S12 and S13) of both the Cox model and GBM was excellent.

4.4 Algorithmic Fairness

AUC stratified by race and sex is given in Appendix 4. Calibration of the Cox model, stratified by race and sex, is shown in Figure 6. Differences in AUC and calibration across race/sex were mostly minimal, but there are notable exceptions. AUC was noisier for

Hispanic individuals than for non-Hispanic individuals, and more noisy for women relative to men. This is attributable to the smaller sample sizes in these groups. In several of the states (although not all), AUC was slightly reduced for Hispanic individuals. For calibration, model predictions were often pessimistic for Hispanic men and women and optimistic for Black men.

Information on how well the models perform on the EEOC's four-fifths rule is shown in Figure 7. Recall that the EEOC has set an informal rule of thumb that disparate impact occurs when the chance of being declared safe to hire for a vulnerable group (e.g. Black applicants) is $< 80\%$ the chance of being declared safe to hire for the nonvulnerable group (e.g. White applicants). We indicate this threshold in the figure, which shows the ratio of the mean predicted risk for each race/sex group relative to the mean predicted risk for non-Hispanic White women. Although we consistently see higher predicted risk for Black and Hispanic individuals than for White individuals, and for men relative to women, these discrepancies do not surpass the EEOC's cutpoint for our Cox and GBM models. However, the use of a 10 year threshold on prior felony convictions violates this guideline for non-Hispanic Black men in Florida, North Carolina, Texas, and Wisconsin.

To better understand the algorithmic bias in our model, we conducted a series of sensitivity analyses in North Carolina. We focus on North Carolina because it showed some challenges with algorithmic bias based on the above results, produced higher risk predictions than the other states, and has some of the best data coverage of the seven states. Those results are given in Appendix 4. We find evidence of lingering racial and sex-based algorithmic bias in the models we developed. In the model that predicted recidivism using the linear predictor of our model along with race and sex, we found that the main effects of and interactions with race and sex were significant.

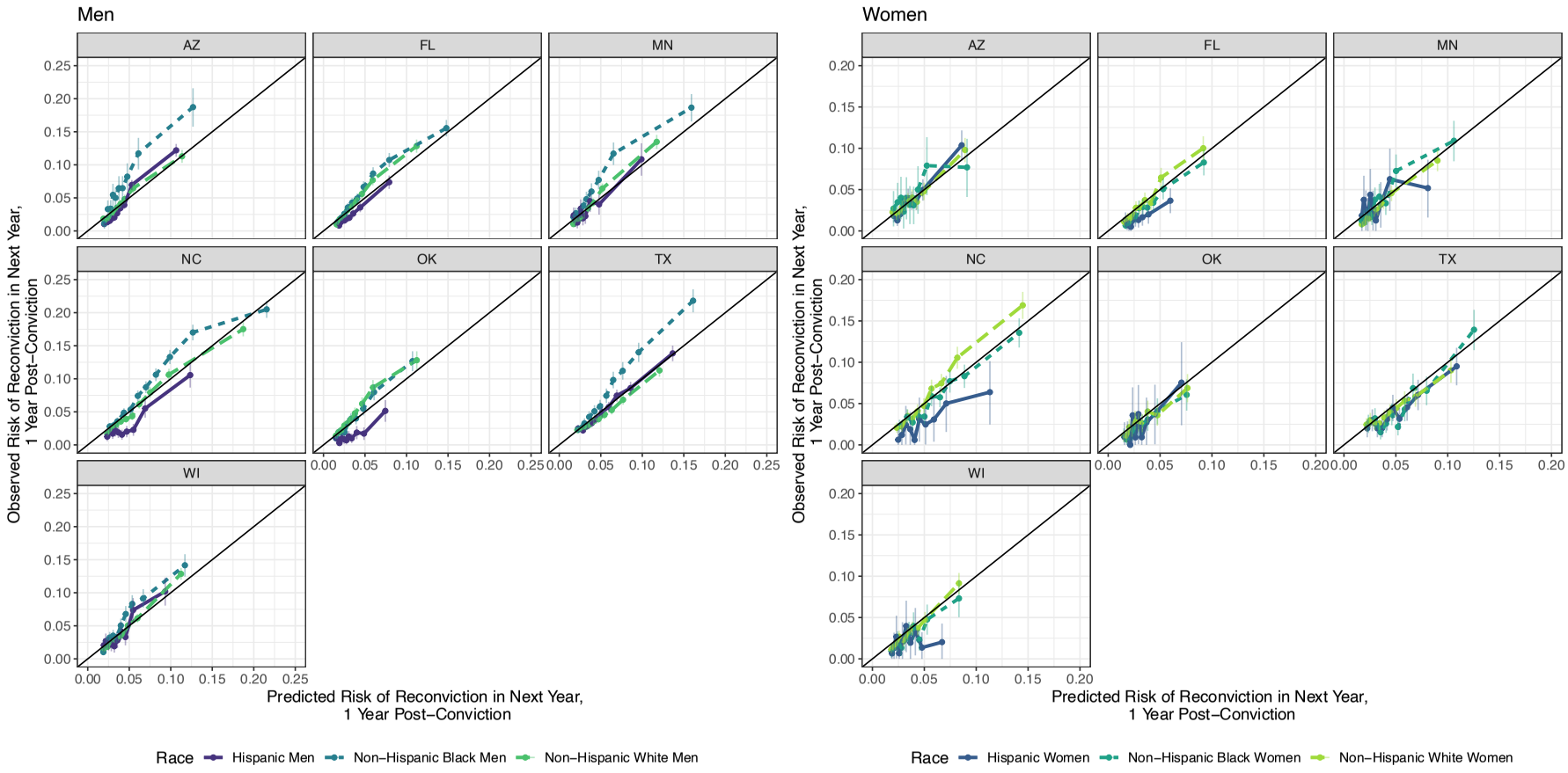


Figure 6: Race- and sex-specific calibration of a Cox proportional hazards model for predicting recidivism, stratified by state. Perfect calibration occurs along the 45 degree black line; groups with calibration further from the 45 degree line are worse calibrated. The top left quadrant indicates that the model generated risk predictions that were lower than they should have been for the group (optimism); the bottom right quadrant indicates that risk predictions were higher than they should have been (pessimism).

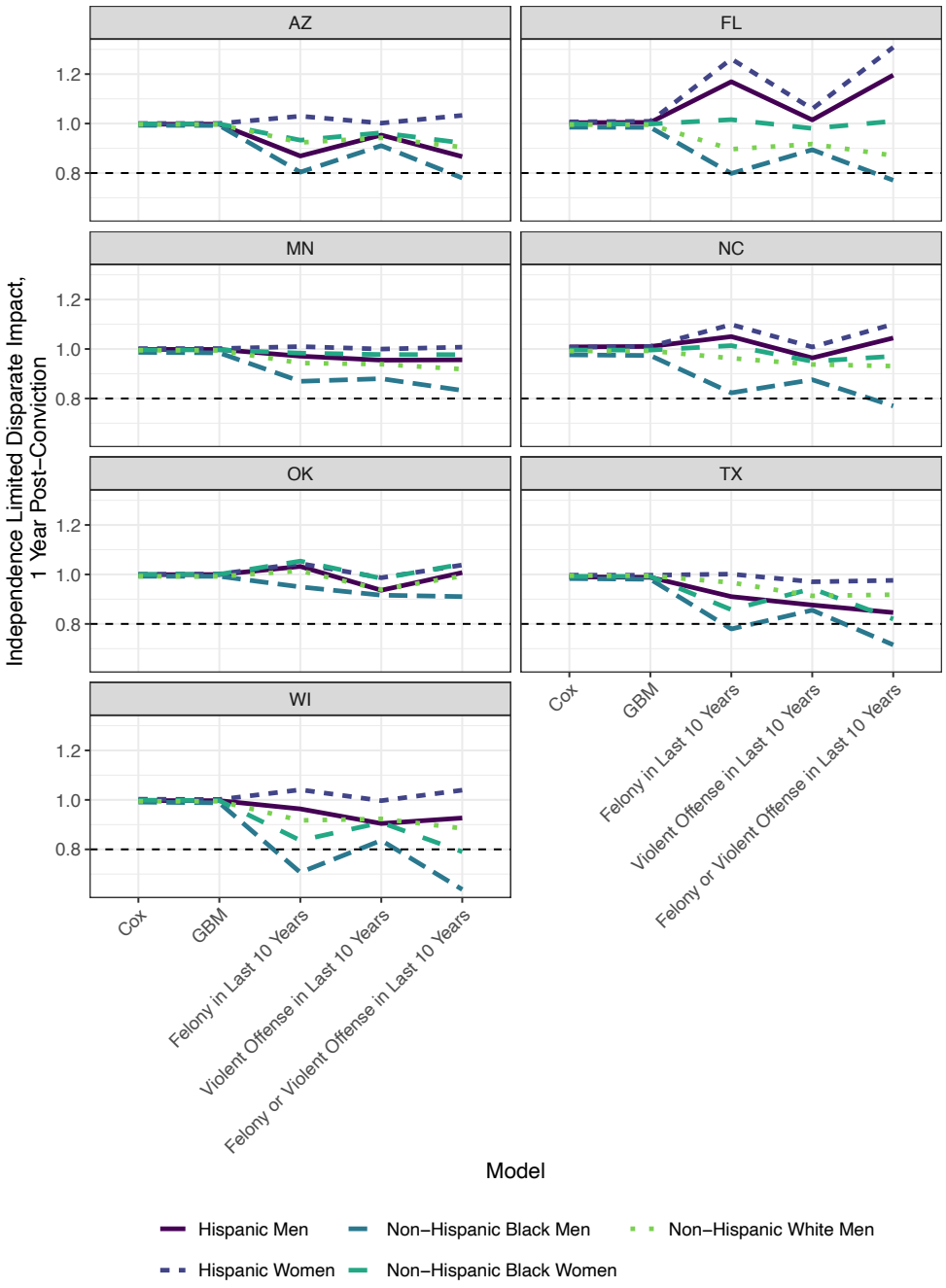


Figure 7: Independence limited disparate impact of recidivism predictions from a Cox proportional hazards model, stratified by state and sex. Models with an independence limited disparate impact falling below the dashed line at 0.8 indicate that for that group, the model fails the Equal Employment Opportunity Commission's rule-of-thumb of no more than 80% disparate impact for protected groups.

When we further included race and sex in the Cox model (both as main effects and interacted with the other criminal history predictors), they continued to be significant predictors, and most of their interactions with other predictors were significant as well, suggesting differential effects (feature bias) of our predictors. When we compare the race/sex-stratified AUC and calibration of the Cox models both with and without race/sex information, we see that the model including race/sex information offers improved predictive performance for Hispanic individuals, non-Hispanic Black men, and non-Hispanic White women—although the improvements in predictive performance for these groups are fairly small.

Taken as a whole, this suggests that criminal history information alone cannot fully account for differences in recidivism rates by race and sex; we see some evidence of feature bias, heterogeneous relationships between predictors and outcomes by race and sex, and different recidivism hazards by race and sex. However, both the Cox model and GBM offer improvements in algorithmic fairness relative to heuristics such as whether or not a person committed a felony in the last 10 years.

4.5 Portability

Appendix 5 presents our results on model portability, with information on the AUC, calibration, and re-ranking of risks when states' models are applied to other states, or when a non-state-stratified model is fit. Here we see some cause for optimism. Although models from different states do not transfer well, a non-stratified model that uses equal samples of data from all states offers decent predictive performance in most states. In some states (Florida, Texas, North Carolina), this non-stratified model provides AUC almost equivalent to the AUC of the state-specific model. Although there are more discrepancies in calibration, these differences are mostly slight and attenuate as time passes from the most

recent conviction. Individual's ranked risks are preserved when a non-state-stratified model is used instead of a state-specific model. Overall, this suggests that a non-state-stratified model using equal samples from each state may be a good option for a national-level model.

4.6 Reduced Model

Although not part of our planned analysis, we fit a reduced Cox model that omitted information on incarceration and crime type. We did so because standard CBCs will not contain information on incarceration or crime type. We anticipated that employers or other hiring decision-makers might be interested in a model without these predictors, and we wanted to assess this reduced model's performance. Appendix 6 presents the coefficient estimates for the reduced model and the AUC and calibration relative to the full Cox model. Race- and sex-stratified calibration of the reduced model is also shown. We find that the reduced model's predictive performance is worse than the full model's performance, but differences are modest. The reduced model's race- and sex-stratified calibration is similar to the full model's, suggesting that dropping incarceration/crime type information has little implication for fairness. Overall, for employers who do not have access to incarceration and/or crime type information, the reduced model would likely be a good option.

4.7 Key Takeaways

To summarize, we want to highlight several key results:

1. **The weighting of multiple criminal records for a given individual has major implications for estimated recidivism rates.** Researchers should think carefully about the population for which they want to estimate recidivism rates, as this should guide the recidivism rate calculation. In our setting (recidivism among job-seekers),

we would recommend sampling a single record proportional to the time spent free in society. Alternative approaches that include all criminal records for a given individual or that sample a single record proportional to the total number of records implicitly upweight records from high-crime periods of an individual's life. In large part due to this finding:

2. **We find that recidivism rates are lower than previous research has found.** More than two-thirds of individuals with a criminal record will not be reconvicted within 10 years of their conviction; in some states this figure is as high as 80%.

3. **We have built a Cox recidivism prediction model with good predictive performance.** Summarizing across states, our model has an AUC of 0.7 for predicting recidivism in the next year one year post-conviction and 0.65 for predicting recidivism in the next year 5 years post-conviction. Calibration is excellent. Predictive performance declines as more time passes since the last conviction.

4. **Our Cox model exhibits some algorithmic bias.** Our model is too pessimistic when predicting recidivism risk for Hispanic individuals, and too optimistic when predicting recidivism risk for Black men. We see evidence of race- and sex-based feature bias for most of the predictors in our model, which likely implies the presence of race- and sex-based label bias as well. However, our model is still an improvement with regards to algorithmic bias relative to a no-felony-in-the-last-10-years rule.

5. **Our Cox model is portable across the seven states we evaluate here.** We find that a Cox model fit using equal samples of data from each state returns adequate predictive performance in all of our seven states.

6. **A reduced Cox model that does not consider incarceration or crime type**

information still offers adequate predictive performance, although slightly worse than the full Cox model. Because employers may not have easy access to incarceration and crime type information, this reduced model may be a good option.

5 Discussion

Our goal in this paper was to develop an accurate, transparent, and fair recidivism risk assessment to assist employers and hiring managers in their decision-making. Overall, we recommend the use of the non-state-stratified Cox proportional hazards model we developed for pre-hiring recidivism risk assessment. We recommend it because based on our review of existing approaches to this problem, we found that our Cox model is more transparent, more accurate, and more fair than existing alternatives. We believe that the most natural place for this model to be used is in the "targeted screening" step of the hiring process. Employers would run CBC results through a risk assessment based on our guidance, which would identify which individuals are safe to hire without further checking, and which individuals should be asked to provide additional information about their criminal record, rehabilitation efforts, and other mitigating factors which would enter into a non-algorithmic individualized assessment of that person's suitability for hire.

Our approach to this problem is groundbreaking in several ways. We are the first team to make use of the Criminal Justice Administrative Records System (CJARS)'s data for this purpose [Finlay and Mueller-Smith, 2022, Finlay et al., 2022]. These new and exciting data allowed us to examine recidivism patterns in seven states from the perspective of employers; with the exception of North Carolina, none of the other six states have ever had their data used to predict recidivism risk in an employment context [Bushway et al., 2022]. The CJARS data allowed us to use convictions (rather than arrests) and to incorporate

information on incarceration, which is important and rarely done [Bushway et al., 2022]. Our approach to cleaning the data prior to model-building improves upon the Bushway et al. [2022] approach to better reflect our primary goal, which is to estimate recidivism risk for potential job applicants. Because an individual who is convicted of multiple crimes in short succession is unlikely to apply for jobs in those short periods, we instead use a time-based sampling approach which gives greater weight to individuals who spent longer periods in the community. Accordingly, we estimated recidivism rates to be lower than have historically been reported. Our careful approach to model-building, predictive performance assessment, and algorithmic bias investigation adds to the literature on risk prediction, particularly in the criminal justice space [Wang et al., 2023]. Finally, our examination of the portability of recidivism models across states sheds new light on how recidivism differs across states [Wang et al., 2023].

We want to highlight some limitations of our model and directions for future work. We think usability and fairness are the chief limitations of our model. With regards to usability, our model still has several limitations that would restrict its actual use. First, our model cannot fix the issues identified in Section 2 with spurious convictions on CBCs or with the confusing/heterogeneous terminology used on CBC results, which may cause issues with translating CBC results into the tidy set of predictors we use here. Second, our model returns a continuous risk prediction ranging from 0-100%. From a statistical perspective, this is preferable, but from the employer's perspective, a binary yes/no decision of whether to clear the applicant for hire or instead flag them for an individualized assessment would likely be preferred. Alternatively, an approach that presents individuals' ranked risk relative to other job candidates may be acceptable. In both cases (binary yes/no or ranked risk), further work is needed on how to implement such an approach and its implications for

predictive performance, algorithmic fairness, and portability across states. Third, we were unable to examine the algorithmic fairness of our model for Asian and Pacific Islander, American Indian and Alaska Native, and "Other race" individuals due to small sample sizes in these groups.

We have argued throughout this paper that our model is an improvement over many common approaches currently in use, but we are not totally satisfied on this point. Further work is needed to evaluate our model relative to more sophisticated comparison approaches than the felony-in-last-10-years heuristic that we considered here. It is challenging to find such comparators because so few of them are made public, but in future work we hope to either obtain some of these competitor models or to develop more realistic options ourselves. Nonetheless, we have provided an illuminating assessment of the predictive performance of a no-felony-in-last-10-years rule, a requirement which is imposed for many jobs by state legislation. Our findings suggest that such a requirement likely runs afoul of the EEOC's guidance on the use of criminal history information in hiring, both because it breaks the four-fifths rule of thumb for Black men and because its predictive performance is poor.

The EEOC has expressed some concern about formulas and algorithms like the one we propose here, particularly when applied broadly. However, in the current hiring landscape, many large employers do use some kind of algorithmic approach for their targeted screen: the question is to what extent statistics was involved in the development of the algorithm. Discouraging the use of statistical modeling may encourage the development of algorithms or rules-of-thumb that reproduce an individual's or society's biases and that may reflect inaccurate assumptions around an individual's recidivism risk over time. That said, we agree that further tailoring of our model would likely be needed, perhaps for particular types of crimes that are of high relevance to the job, and such a model would need to be

carefully checked and recalibrated to ensure sustained accuracy, fairness, and rigor over time. As improved algorithmic approaches and data sources are developed, policy-makers should carefully consider whether encouraging employers to rely on human decision-making in this area actually produces better outcomes, or whether there could be benefits from using a system in which fairness can be more rigorously assessed and optimized.

With regards to fairness, our models show some algorithmic bias. Based on our sensitivity analyses, this algorithmic bias is somewhat across-the-board, affecting many of our predictors in a way that suggests both feature and label bias in this setting. While some of these problems may be resolved with better data, we think the algorithmic bias that we observe may reflect the stark reality of how race operates in the United States. Black Americans are at substantially higher risk of being convicted of crimes [Neil and Zanger-Tishler, 2025], a situation that reflects historic inequities in wealth, access to education and employment, neighborhood safety and health, policing, and interactions with the criminal justice system. These societal problems translate into feature bias and label bias in our data, making it challenging to accurately assess recidivism risk while restricted to the EEOC's list of acceptable predictors. It is striking that our model actually returns *optimistic* predictions of recidivism risk for Black men, highlighting that Black men are reconvicted more quickly than their age and criminal history would suggest, likely due to the social factors we have just described. In keeping with theoretical findings by Corbett-Davies et al. [2023], we were able to reduce some of the algorithmic bias in prediction accuracy and calibration by adjusting for race and sex in our model, because including these predictors captures some of the confounding effects (wealth, employment, neighborhood, policing) that we listed. But including predictors like race or sex in a model for employment decision-making would be obviously unacceptable. There are other approaches for stripping out some of the al-

gorithmic bias we observe (or constraining it), but as Wang et al. [2023] note, many of these approaches make the resulting model nontransparent, and may still violate legal requirements on the use of race/sex in employment decisions.

On this topic, we echo the recommendations of Corbett-Davies et al. [2023], which is that the best solution may be for policymakers to detach risk assessment from risk-informed decision-making. Statisticians can develop optimal risk assessments with respect to accuracy and can quantify the differential impact of those risk assessments. In turn, policymakers could agree on how to use those risk predictions: perhaps differently for different groups, in light of historic inequities, or to achieve particular social priorities. Despite the lingering algorithmic bias of our model, we would still recommend the use of the model we developed here, because its algorithmic bias appears reduced relative to alternatives currently in use, and *all* decision-makers in this space (human or algorithmic) would be vulnerable to the challenges with feature and label bias that we document here.

We envision several avenues for future statistical work in this space. Here, we used simple survival models to predict recidivism, but there are more sophisticated alternatives. Prior research has suggested that cure models can produce better fits for recidivism, as some individuals do seem to be "cured" of criminal behavior [Bushway and Kalra, 2021]; this may be one path to improve upon our current models. Recurrent-event models may offer an elegant approach for using an individual's complete criminal record trajectory, rather than having to sample a single criminal episode as we do here. Time-dependent covariates may be another fruitful direction; we saw some evidence that the use of a time-dependent covariate to model time since last offense might be appropriate. More broadly, if a dataset with information on rehabilitation activities or employment history post-conviction became available, these could also be incorporated as time-dependent covariates, although there

may be challenges with doing so within the context of a prediction model. It also may be beneficial to treat this problem as a logistic regression for risk of reconviction within the next year, fit in successive cohorts of time since last offense. This would be conceptually similar to what we do here, but may enable the use of approaches like Liu et al. [2022] for generating transparent risk prediction models. Finally, it may be interesting to explore the consequences of using conviction as a surrogate for job performance, which is effectively what employers are doing. Although the EEOC recognizes recidivism risk as a valid concern unto itself for employers to consider [Equal Employment Opportunity Commission, 2012], there is no documented relationship between recidivism risk and job performance [Bushway and Kalra, 2021]. Data on this topic are limited, but if they became available, using statistical methodology to evaluate reconviction as a surrogate for job performance may yield interesting results.

SUPPLEMENTARY MATERIAL

Supplementary Materials: More information on data cleaning and additional model performance results. (PDF)

Disclosure Statement

The authors report there are no competing interests to declare.

References

SD Bushway, BG Vegetabile, N Kalra, L Remi, and G Baumann. Providing another chance: Resetting recidivism risk in criminal background checks. Research Report RRA1360-1, RAND, 2022.

KM Kuhn. Is it disqualifying? Practitioner responses to criminal offenses in hiring decisions. *Equality, Diversity and Inclusion: An International Journal*, 38(5):547–563, 2019.

SK Baier. Reducing employment barriers for people with criminal records. *Journal of Corporation Law*, 46(1):219–246, 2020.

MA Stoll and SD Bushway. The effect of criminal background checks on hiring ex-offenders. *Criminology & Public Policy*, 7(3):371–404, 2008.

D Pager. The mark of a criminal record. *American Journal of Sociology*, 108(5):937–975, 2003.

A Agan and S Starr. The effect of criminal records on access to employment. *American Economic Review: Papers & Proceedings*, 107(5):560–564, 2017.

FJ Cavico, BG Mujtaba, and SC Muffler. Criminal background checks in employment: An unfolding legal quandary for employers. *Journal of Law and Criminal Justice*, 2(1): 41–103, March 2014.

J Levashina, JA Peck, and L Ficht. Don't select until you check: Expected background checking practices. *Employer Responsibilities and Rights Journal*, 29:127–148, 2017.

SD Bushway and JT Pickett. Direct incentives may increase employment of people with criminal records. *Criminology & Public Policy*, pages https://doi.org/10.1111/1745–9133.12681, 2024.

Z Cullen, W Dobbie, and M Hoffman. Increasing the demand for workers with a criminal record. *The Quarterly Journal of Economics*, 138(1):103–150, 2023.

L Couloute and D Kopf. Out of prison & out of work: Unemployment among formerly incarcerated people. https://www.prisonpolicy.org/reports/outofwork.html, July 2018.

M Denver, G Siwach, and SD Bushway. A new look at the employment and recidivism relationship through the lens of a criminal background check. *Criminology*, 55(1):174–204, 2017.

S Ferguson. Understanding America's labor shortage. https://www.uschamber.com/workforce/understanding-americas-labor-shortage, June 2024.

K Finlay and M Mueller-Smith. Criminal Justice Administrative Records System (CJARS) [dataset]. https://cjars.isr.umich.edu., 2022.

K Finlay, M Mueller-Smith, and J Papp. The Criminal Justice Administrative Records System: a next-generation research data platform. *Scientific Data*, 9(562):197–228, 2022.

SD Bushway and N Kalra. A policy review of employers' open access to conviction records. *Annual Review of Criminology*, 4:165–189, 2021.

A Blumstein and K Nakamura. Redemption in the presence of widespread criminal background checks. *Criminology*, 47(2):327–359, 2009.

JJ Smith. Banning the box but keeping the discrimination? Disparate impact and employers' overreliance on criminal background checks. *Harvard Civil Rights - Civil Liberties Law Review*, 49(1):197–228, 2014.

M Kurlychek, SD Bushway, and M Denver. Defensible decisions: Balancing employer and prospective employee rights in an era of criminal background checks. *Equality, Diversity and Inclusion: An International Journal*, 5(38):529–546, 2019.

D McElhattan. The proliferation of criminal background check laws in the United States. *American Journal of Sociology*, 127(4):1037–1093, January 2022.

Equal Employment Opportunity Commission. Title VII, 29 CFR Part 1601, 29 CFR Part 1606, 29 CFR Part 1607. https://www.eeoc.gov/laws/guidance/enforcement-guidance-consideration-arrest-and-conviction-records-employment-decisions, 2012.

R Neil and M Zanger-Tishler. Algorithmic bias in criminal risk assessment: The consequences of racial differences in arrest as a measure of crime. *Annual Review of Criminology*, 8(4):1–23, 2025.

Equal Employment Opportunity Commission. Title VII, 29 CFR Part 1601. https://www.eeoc.gov/laws/guidance/what-you-should-know-eeocs-response-letter-state-attorneys-general-use-criminal: :text=First

Jacqueline Ramos, et al., v. Walmart, Inc. Case No. 2:21-cv-13827-CCC-AME in the U.S. District Court for the District of New Jersey. https://www.outtengolden.com/insights/media/news/outten-golden-llp-and-youth-represent-criminal-history-hiring-discrimination-lawsuit-against-walmart-is-expanded/, 2021.

Equal Employment Opportunity Commission. Title VII, 29 CFR Part 1607. https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial, 2023.

APTMetrics. Criminal background check evaluations. https://aptmetrics.com/background-check/, 2024.

VeriFirst Background Screening, LLC. Background screening decision matrix. https://screen.verifirst.com/eeoc-background-screening-decision-matrix-cos, 2024.

Cisive. What happens when a candidate background check reveals a criminal

record? https://blog.cisive.com/what-happens-when-a-candidate-background-check-reveals-a-criminal-record, 2021.

Yardstik. Background check adjudication: Definition, guide, and best practices. https://yardstik.com/blog/background-check-adjudication/, 2024.

MR Durose, HN Snyder, and AD Cooper. Multistate criminal history patterns of prisoners released in 30 states. Technical report, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, Washington, DC, 2015.

DR Cox. Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, January 1972.

HA Chipman, EI George, and RE McCulloch. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

RA Sparapani, BR Logan, RE McCulloch, and PW Laud. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753, July 2016.

J Liu, C Zhong, B Li, M Seltzer, and C Rudin. Fasterrisk: Fast and accurate interpretable risk scores. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

RA Sparapani, C Spanbauer, and R McCulloch. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1), January 2021.

G Ridgeway. Generalized boosted models: A guide to the gbm package. https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf, June 2024.

P Blanche, JF Dartigues, and H Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, December 2013.

I Cabreros, J Snoke, OA Osoba, I Khan, and MN Elliott. Advancing equitable decision-making for the Department of Defense through fairness in machine learning. Technical report, RAND Corporation, Santa Monica, CA, 2023.

S Corbett-Davies, JD Gaebler, H Nilforoshan, R Shroff, and S Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24:1–117, 2023.

B Grunwald. Racial bias in criminal records. *Journal of Quantitative Criminology*, 40: 489–531, 2023.

C Wang, B Han, B Patel, and C Rudin. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 39:519–581, 2023.

Supplementary Materials for "Safe to Hire: Predicting Recidivism Risk for Job Candidates with Criminal Records"

Elizabeth C. Chase, Shawn D. Bushway,
Bethany Saunders-Medina, Lane F. Burgette

January 17, 2025

Contents

1 Appendix 1: More Information on Data Cleaning **2**

2 Appendix 2: Sample Construction Flow Charts **8**

3 Appendix 3: Additional Results on Predictive Performance **15**

4 Appendix 4: Additional Results on Algorithmic Fairness **19**

5 Appendix 5: Results on Model Portability **25**

6 Appendix 6: Results on Reduced Model **30**

1 Appendix 1: More Information on Data Cleaning

The Criminal Justice Administrative Records System (CJARS) data are maintained in three sets: the adjudication set (which contains information on all adjudications, which may or may not have resulted in conviction), the incarceration set (which contains information on incarcerations), and the demographic set (which contains demographic details on every person appearing in any CJARS dataset). These three sets are linked across each other by a common individual-specific identifier.

Data cleaning was carried out separately on each state. On each state, we did as follows:

1. In the demographic set, there are some individuals who appear multiple times but with differing race/sex/date of birth (DOB) information. Following CJARS guidance, when this occurred we took the modal race/sex/DOB to collapse to a single demographic record. When there were multiple modes, we randomly selected among the modes.

2. We restricted the adjudication set to only individuals who had at least one adjudication in the state of interest (either AZ, FL, MN, NC, OK, TX, WI).

3. We removed adjudications that resulted in the death penalty.

4. We restricted to convictions: we removed adjudications that resulted in acquittal, not-guilty, or dismissed charges.

5. We removed convictions that were missing date information.

6. Convictions were grouped into criminal episodes using a 30-day threshold, e.g. a cluster of convictions that all occurred in a single 30-day span was assumed to be part of a single crime episode. Note that this 30-day threshold was not rolling. Table S1 gives information on how many convictions were usually grouped under this threshold.

7. Individuals whose DOB implied that they were younger than 15 or older than 100 at the time of their criminal episode had their DOB set to missing.

8. On the basis of the cleaned conviction set, we calculated conviction histories: number of prior offenses in various crime categories at the time of each conviction. We also calculated age at each conviction and age at first conviction.

9. We restricted the incarceration set to only include individuals who also appeared in the cleaned conviction set.

10. We removed incarcerations that were missing an entry date or that occurred prior to a designated start date. (For most states, this start date limit was 1900; in NC, due to changes in sentencing guidelines and incarceration record-keeping, we only kept incarcerations that started in 1996 onwards.)

11. Each incarceration was matched to the most recent conviction for a given individual. Note that in some cases multiple incarcerations were assigned to the same conviction, usually indicating prison transfer or the start of another sentence. Also note that we did not limit how much time could pass between the most recent conviction and the incarceration. Tables S2 and S3 give some information on these issues.

12. Demographic (race, sex, age at conviction) and conviction characteristics (minimum sentencing time, number of felony convictions, number of misdemeanor convictions, number of prior felonies, number of prior misdemeanors) were used within a Cox model to predict incarceration duration. A second Cox model that did not include demographic information was fit, needed for individuals with missing demographic information. The median predicted incarceration duration from these models was used to deterministically impute incarceration exit date for incarcerations that were missing an exit date (and that were not marked as ongoing by the CJARS records).

13. Convictions that occurred during incarceration were removed from the dataset. These convictions were retained as part of the individual's conviction record, but were not considered for sampling into model training/validation. Incarcerations were then re-matched to the most recent conviction.

14. Convictions were restricted to a specified date range, given in Table 1 of the main manuscript. (Convictions falling outside of the date range were retained as part of the individual's conviction record, but were not considered for sampling into model training/validation.)

15. Convictions were restricted to individuals ages 15-60 at the time of conviction.(Convictions falling outside of the age range were retained as part of the individual's conviction record, but were not considered for sampling into model training/validation.)

Table S1: Information on the number of convictions grouped into each 30-day thresholded crime episode, stratified by state, using data from the Criminal Justice Administrative Records System (CJARS).

# Convictions	AZ	FL	MN	NC	OK	TX	WI
1	79.9%	76.3%	93.5%	78.0%	74.9%	83.1%	82.4%
2	16.1%	15.5%	5.6%	14.6%	16.0%	14.2%	13.3%
3	2.9%	4.9%	0.7%	4.2%	5.5%	2.1%	3.2%
4	0.7%	1.9%	0.1%	1.8%	2.1%	0.4%	0.8%
5	0.2%	0.8%	0.0%	0.8%	0.8%	0.1%	0.2%
6+	0.1%	0.6%	0.0%	0.6%	0.6%	0.0%	0.1%

16. A single conviction was sampled per individual, with sampling done proportional to the individual's time spent free in society after each conviction. More information on this choice is given in Figures S1 and S2 and is discussed in the main manuscript.

17. This final dataset is regarded as the cleaned dataset for that state. For each state, records were randomly assigned to training and validation sets.

Table S1 gives information on the number of convictions that were grouped into a single crime episode under our 30-day non-rolling threshold. For roughly three-quarters of individuals across states, their criminal episode was comprised of only a single conviction. An additional 15% had two convictions making up the crime episode. However, for small minorities of individuals, a crime episode could involve 3+ convictions. For the most part, patterns were fairly stable across states, with the notable exception of Minnesota, for which 94% of individuals had a single conviction crime episode.

Table S2 gives information on how many incarcerations were grouped per crime episode. Of crime episodes that resulted in incarceration time, around 90% involved only a single incarceration. Note that there is more heterogeneity in the distribution of number of incarcerations across states than there was for the number of convictions, perhaps due to differences in state sentencing approaches, prison capacity, and incarceration systems.

Table S3 shows the distribution of how much time lapsed between the most recent conviction and the start of incarceration (among individuals who were incarcerated for their offense). We would normally expect relatively little time to lapse between conviction

Table S2: Information on the number of incarcerations grouped per crime episode (among crime episodes that resulted in some incarceration), stratified by state, using data from the Criminal Justice Administrative Records System (CJARS).

# Incarcerations	AZ	FL	MN	NC	OK	TX	WI
1	88.9%	94.5%	79.7%	94.8%	94.5%	93.4%	76.7%
2	10.3%	3.7%	14.5%	4.6%	4.1%	4.4%	16.5%
3	0.6%	1.0%	3.7%	0.4%	0.8%	1.2%	4.8%
4+	0.1%	0.9%	2.0%	0.2%	0.6%	1.0%	2.0%

and the start of incarceration. For some states that is the case, but there is large state-to-state variability in how much time lapses between conviction and incarceration. We have some hypotheses about this issue, although none are fully satisfactory. First, CJARS does not contain information on county jails, and individuals may be being held in county jails for some time prior to transfer to the state prison system. Second, individuals may be being convicted in one state and incarcerated in a different state, and if either the conviction or incarceration does not occur in the CJARS universe, we may be missing it from our data, which could also explain the long lags.

Figure S1 shows the distribution of number of criminal episodes per individual. The vast majority of offenders have only a single criminal episode during our follow-up period. However, there is a long right tail of individuals who have many (15+) criminal episodes during our follow-up. Adjusting for these individuals is a critical challenge, because including all of their criminal episodes in our analysis would upweight these individuals relative to less-frequent offenders.

We considered several appraoches for addressing the issue of multiple criminal episodes per individual. In previous work, Bushway et al. [2022] randomly sampled a single criminal episode for each individual, by sampling proportional to the number of records (e.g. if a person had 3 criminal episodes, each episode would have probability 1/3 of being selected). However, our data show that individuals go through "high crime" life phases (usually in their youth) during which time they may have several crime episodes in rapid succession, but then have long crime-free stretches after that point. Randomly sampling a single record proportional to the number of records ignores this issue and increases the chances that the sampled record is from a high-crime period (thus making recidivism rates look higher than

Table S3: Information on the length of time between most recent conviction and incarceration, stratified by state, using data from the Criminal Justice Administrative Records System (CJARS).

Time Lag	AZ	FL	MN	NC	OK	TX	WI
Same day	19.1%	1.1%	23.1%	26.8%	2.1%	0.3%	0.9%
1-7 days	0.9%	16.2%	31.0%	30.7%	4.4%	6.0%	24.3%
8-14 days	2.9%	30.1%	5.3%	18.6%	11.5%	8.2%	14.8%
15-31 days	20.3%	35.0%	4.2%	15.1%	24.1%	23.2%	12.5%
32-90 days	32.2%	13.0%	8.7%	5.3%	36.4%	42.7%	17.7%
3-6 months	4.4%	1.3%	5.6%	0.6%	11.3%	8.7%	6.6%
6-12 months	5.0%	0.7%	7.5%	1.0%	5.8%	5.9%	6.2%
1+ years	15.2%	2.5%	14.6%	1.9%	4.4%	5.0%	16.9%

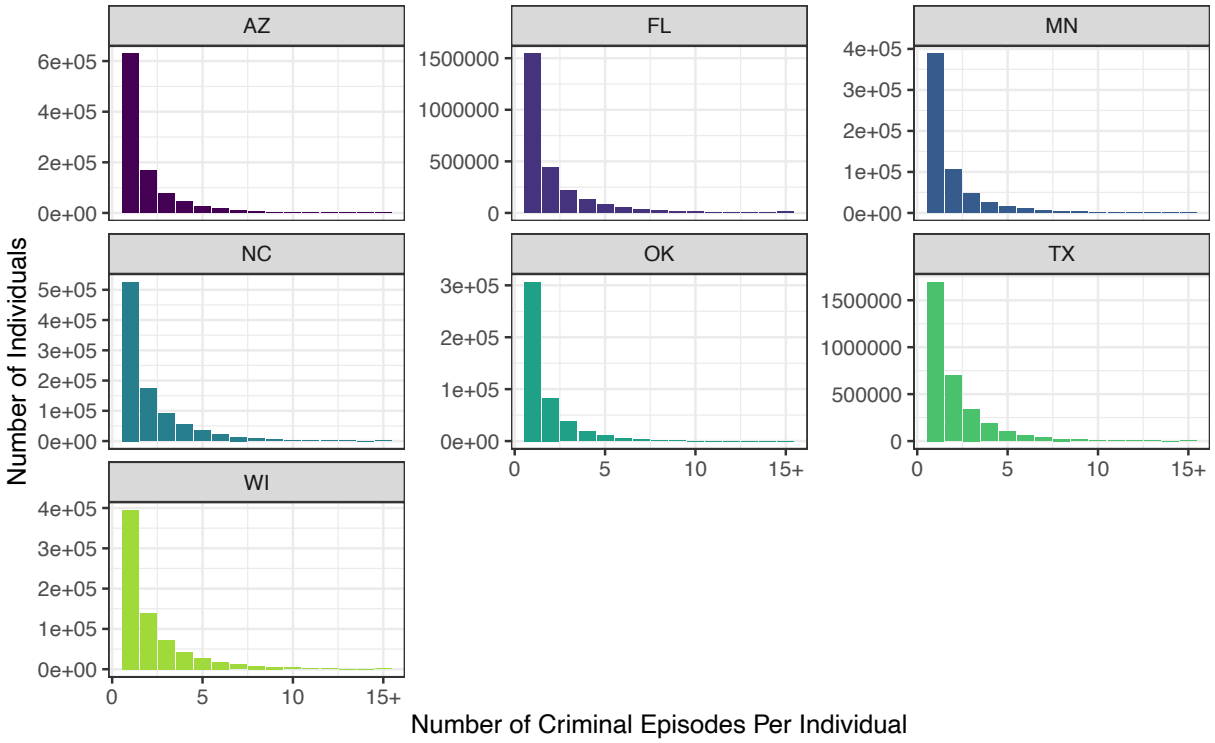


Figure S1: Number of conviction criminal episodes observed per individual in a sample of $n = 9,509,250$ individuals from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

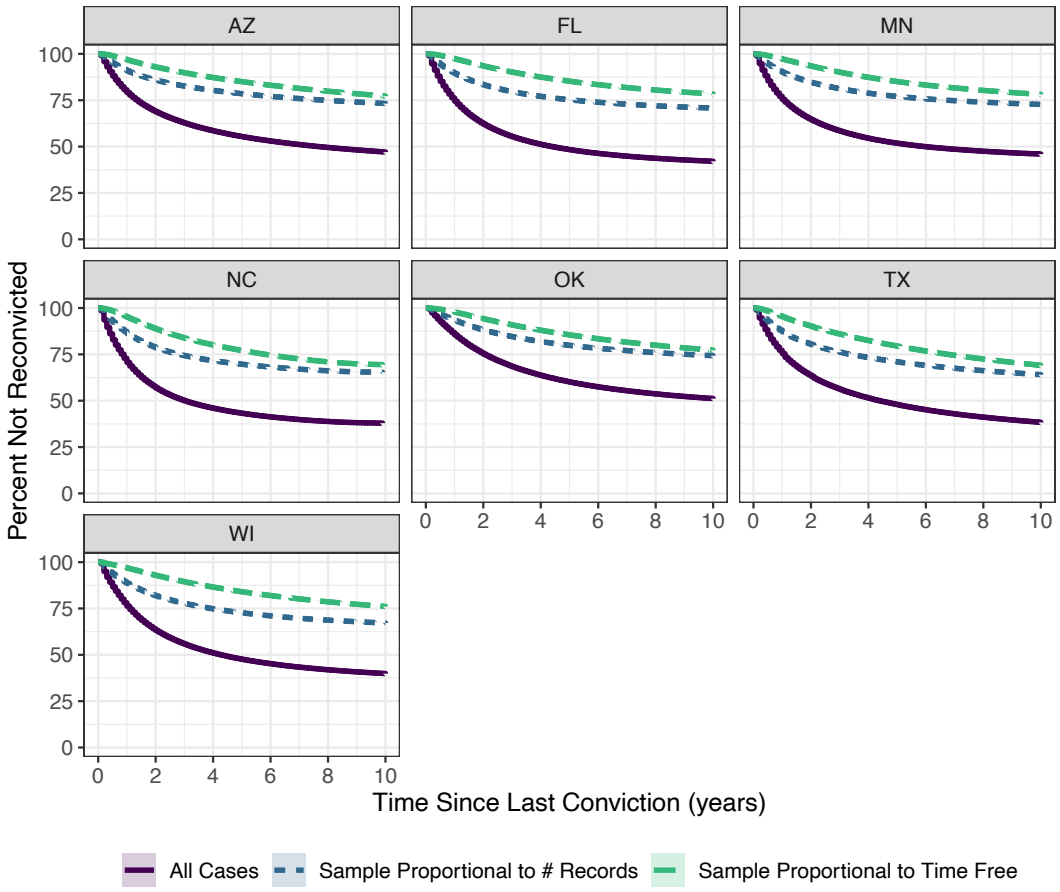


Figure S2: Estimated recidivism rate under three different approaches for sampling criminal episode records from among multiple records for a given individual, stratified by state, in a sample of $n = 9,509,250$ individuals from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

they actually are). Instead, we propose to sample a record proportional to how much time was spent free in society after that episode, which we think better reflects recidivism risk among job candidates—as most job candidates will be applying to jobs during a lower-crime period. As Figure S2 demonstrates, the choice of sampling approach makes a sizable difference to estimated recidivism rates.

2 Appendix 2: Sample Construction Flow Charts

Here are flow charts indicating sample exclusions made when constructing each state's cleaned dataset.

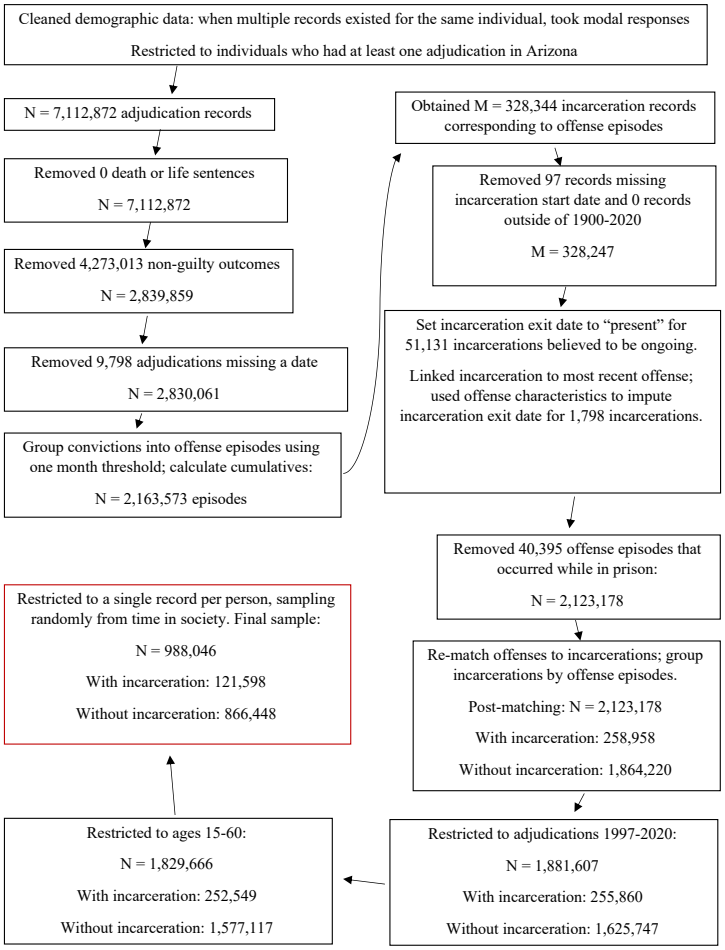


Figure S3: Sample exclusions to construct the cleaned Arizona dataset.

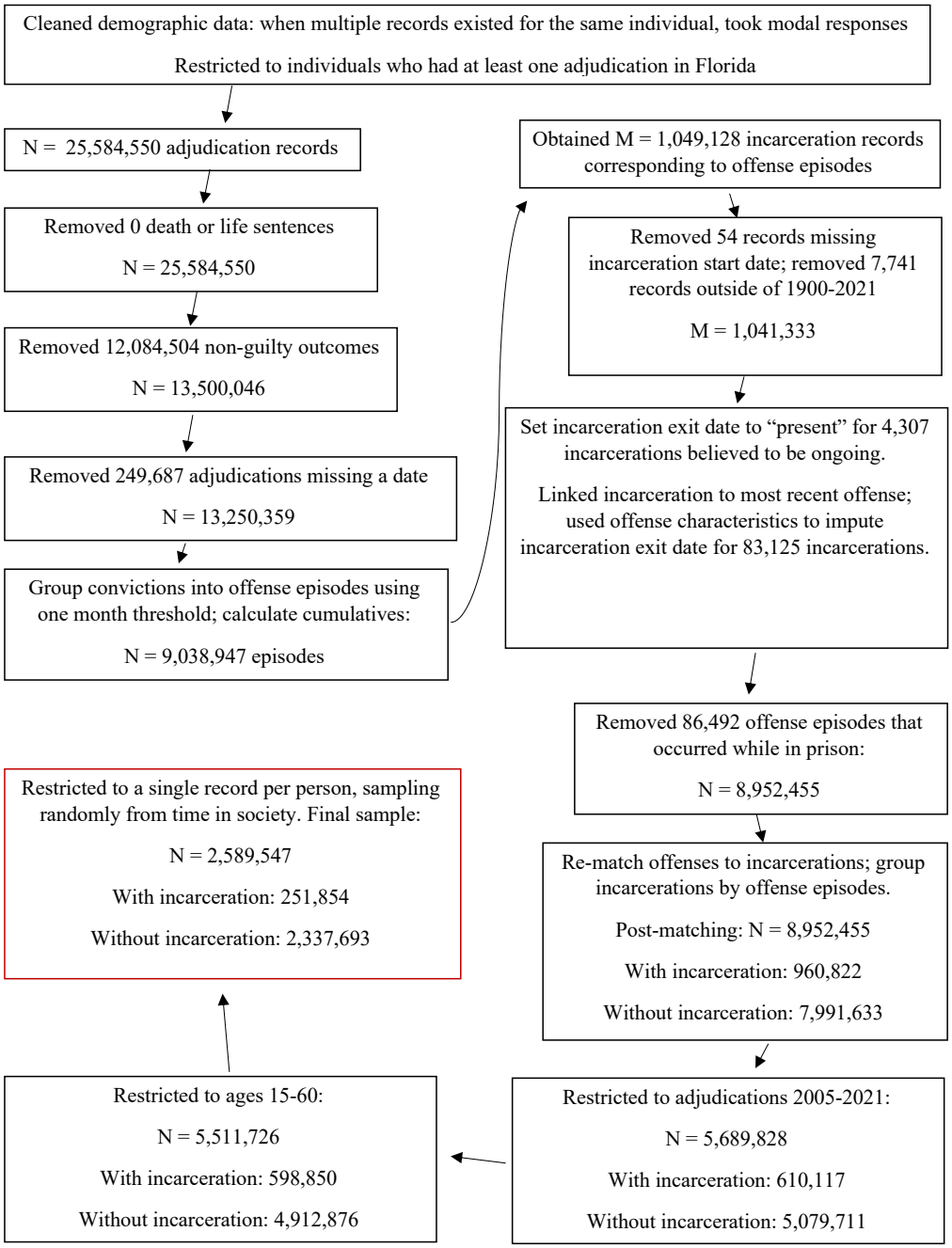


Figure S4: Sample exclusions to construct the cleaned Florida dataset.

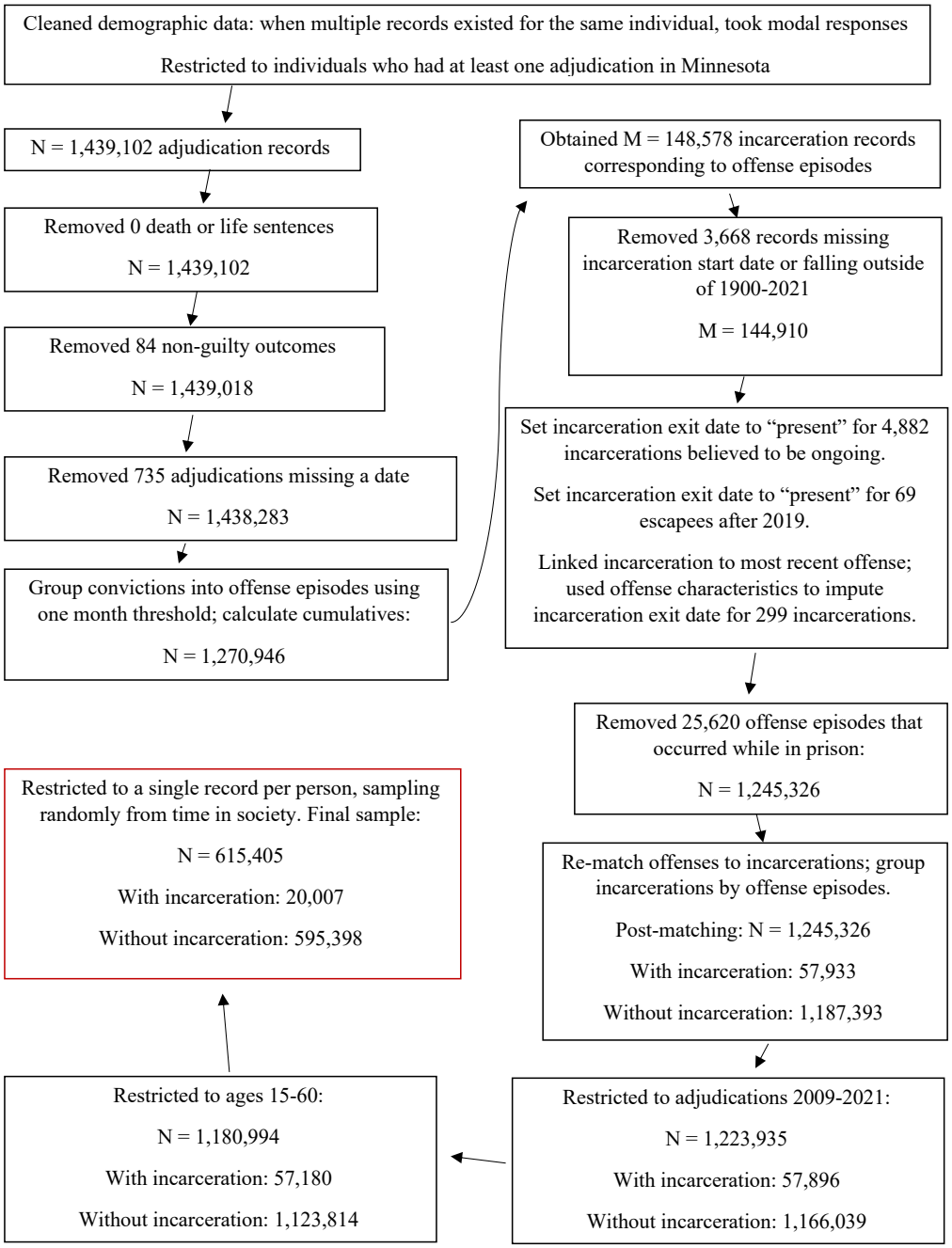


Figure S5: Sample exclusions to construct the cleaned Minnesota dataset.

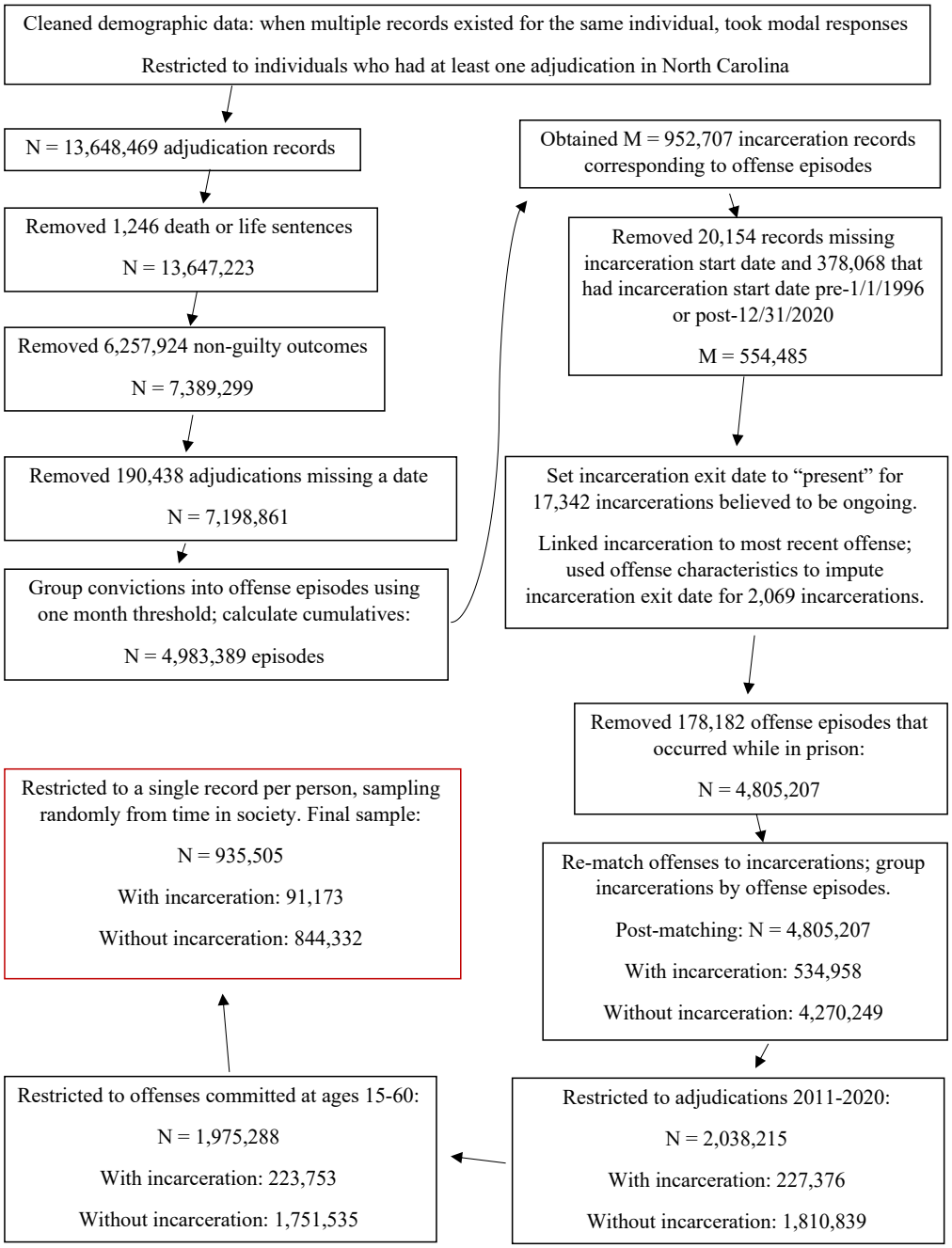


Figure S6: Sample exclusions to construct the cleaned North Carolina dataset.

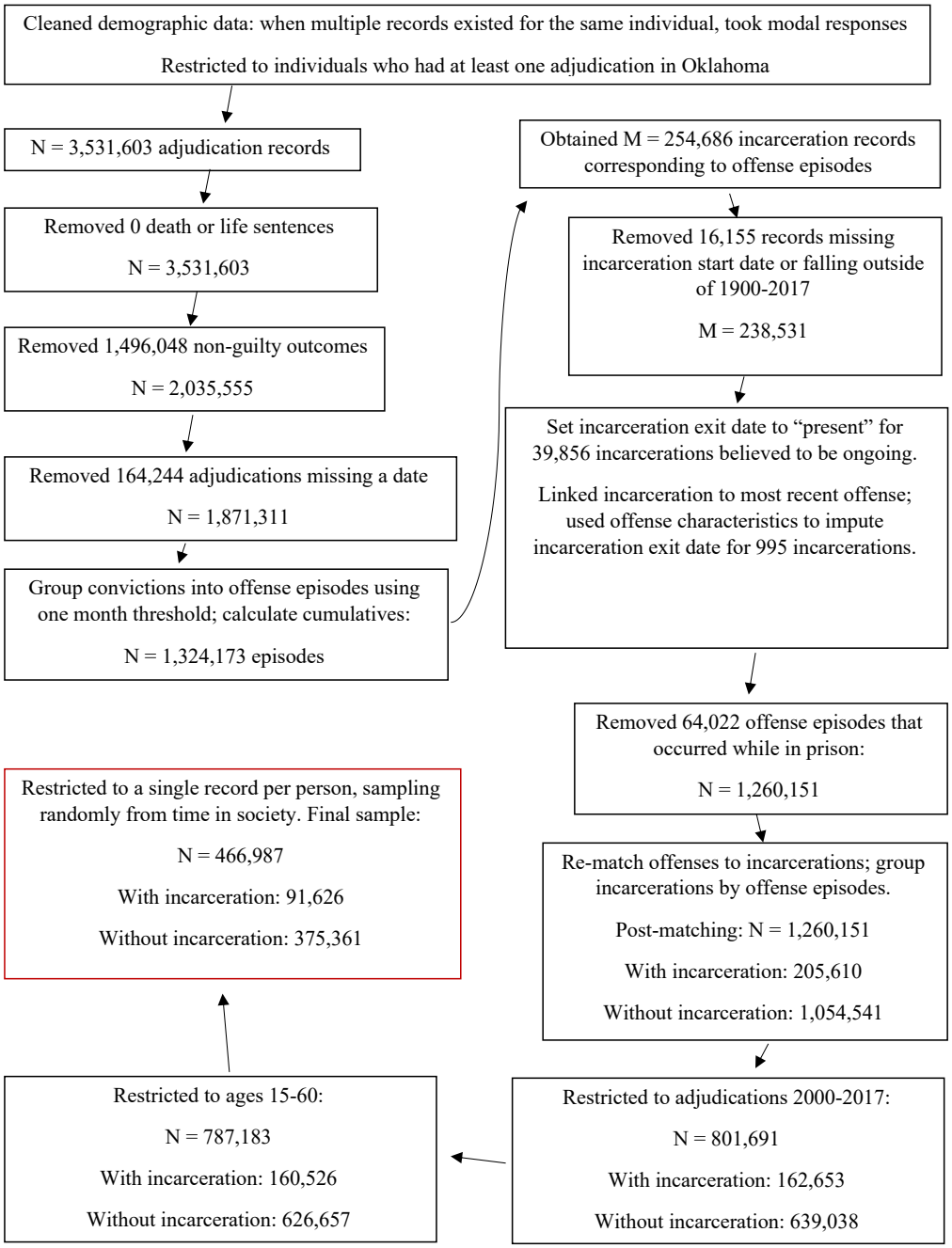


Figure S7: Sample exclusions to construct the cleaned Oklahoma dataset.

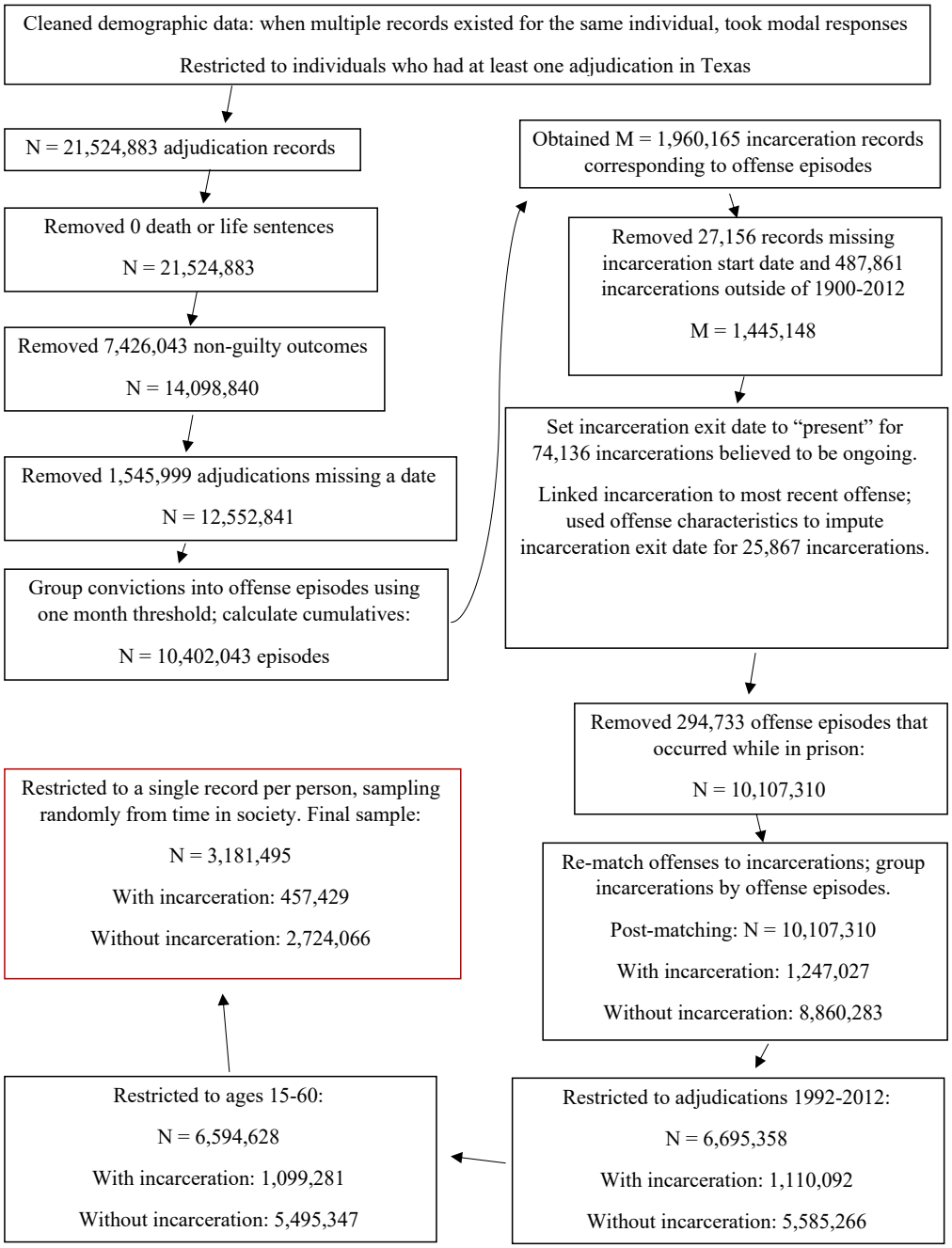


Figure S8: Sample exclusions to construct the cleaned Texas dataset.

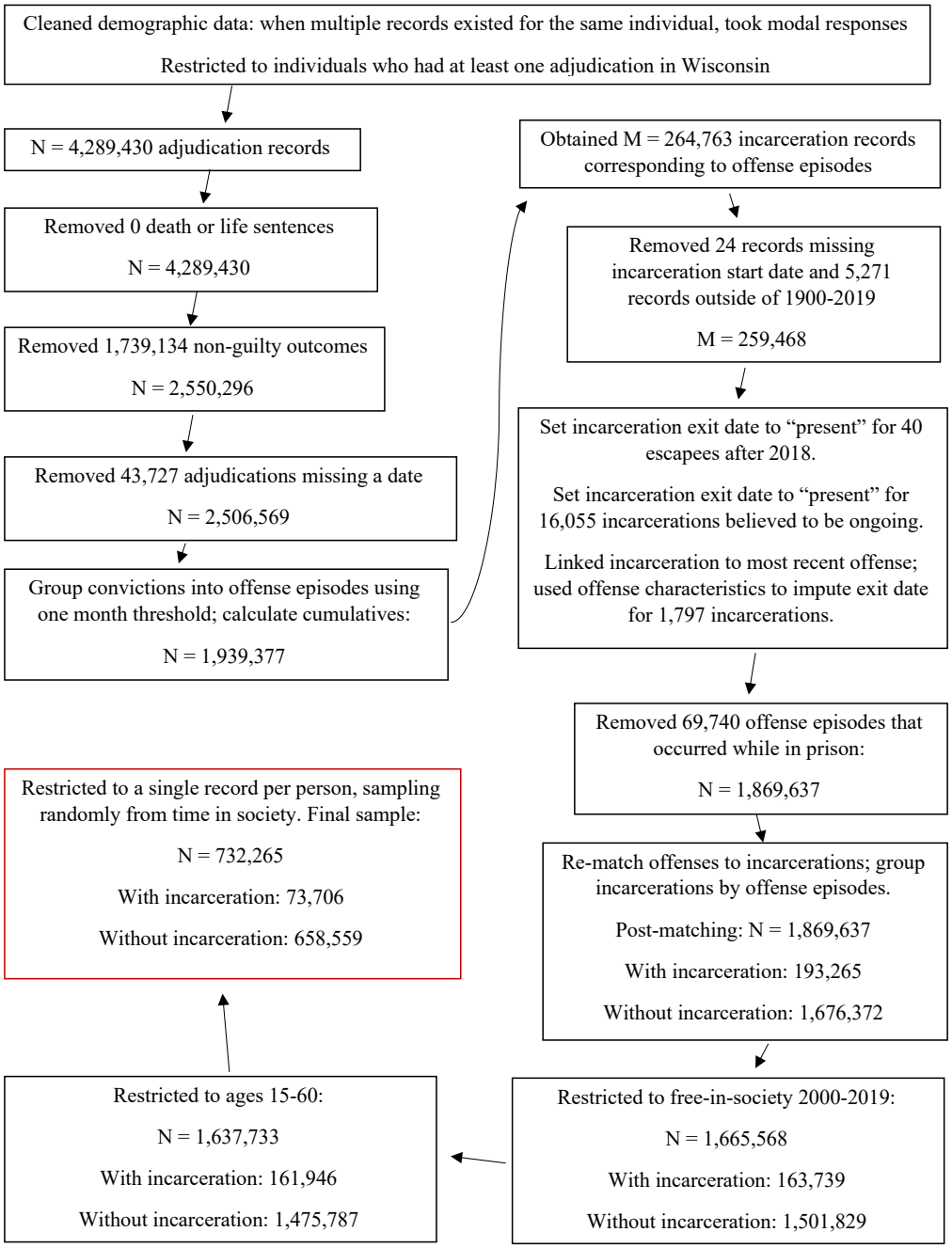


Figure S9: Sample exclusions to construct the cleaned Wisconsin dataset.

3 Appendix 3: Additional Results on Predictive Performance

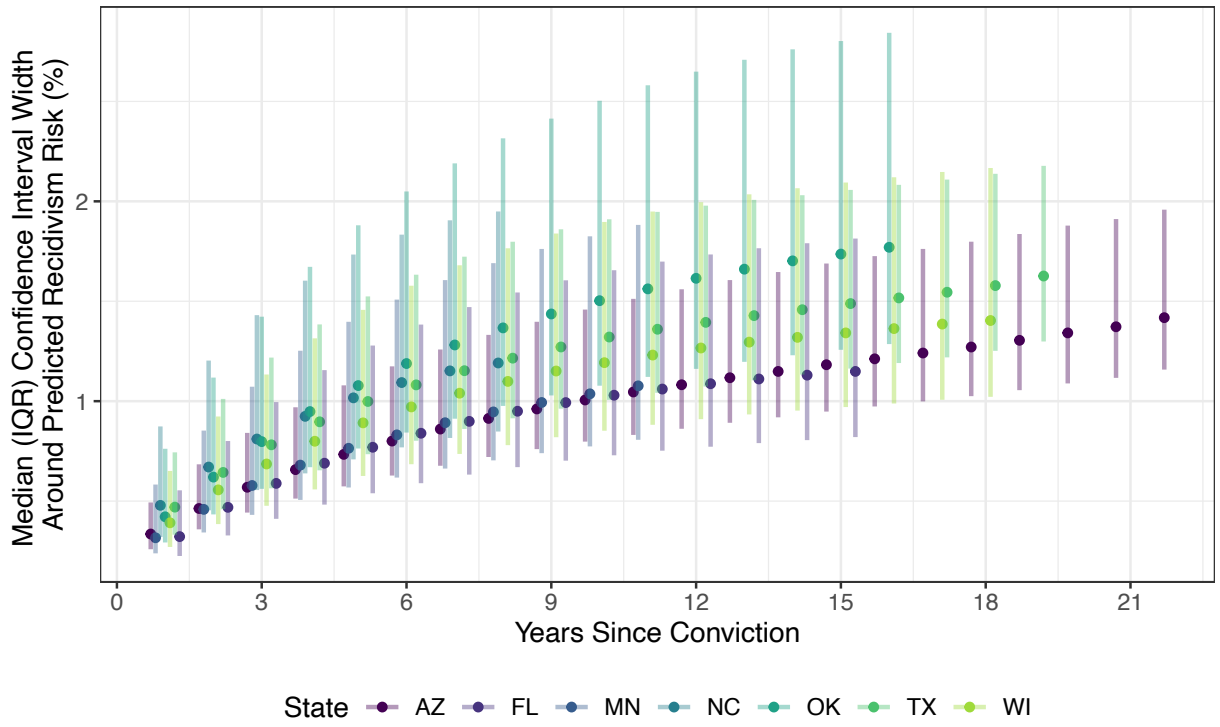


Figure S10: Median and interquartile range for the confidence interval width around predicted recidivism risk, as estimated by a Cox model, stratified by state. Predictions generated on a validation sample of criminal records from the Criminal Justice Administrative Records System data (CJARS). Data range from 1992-2021, depending on the state.

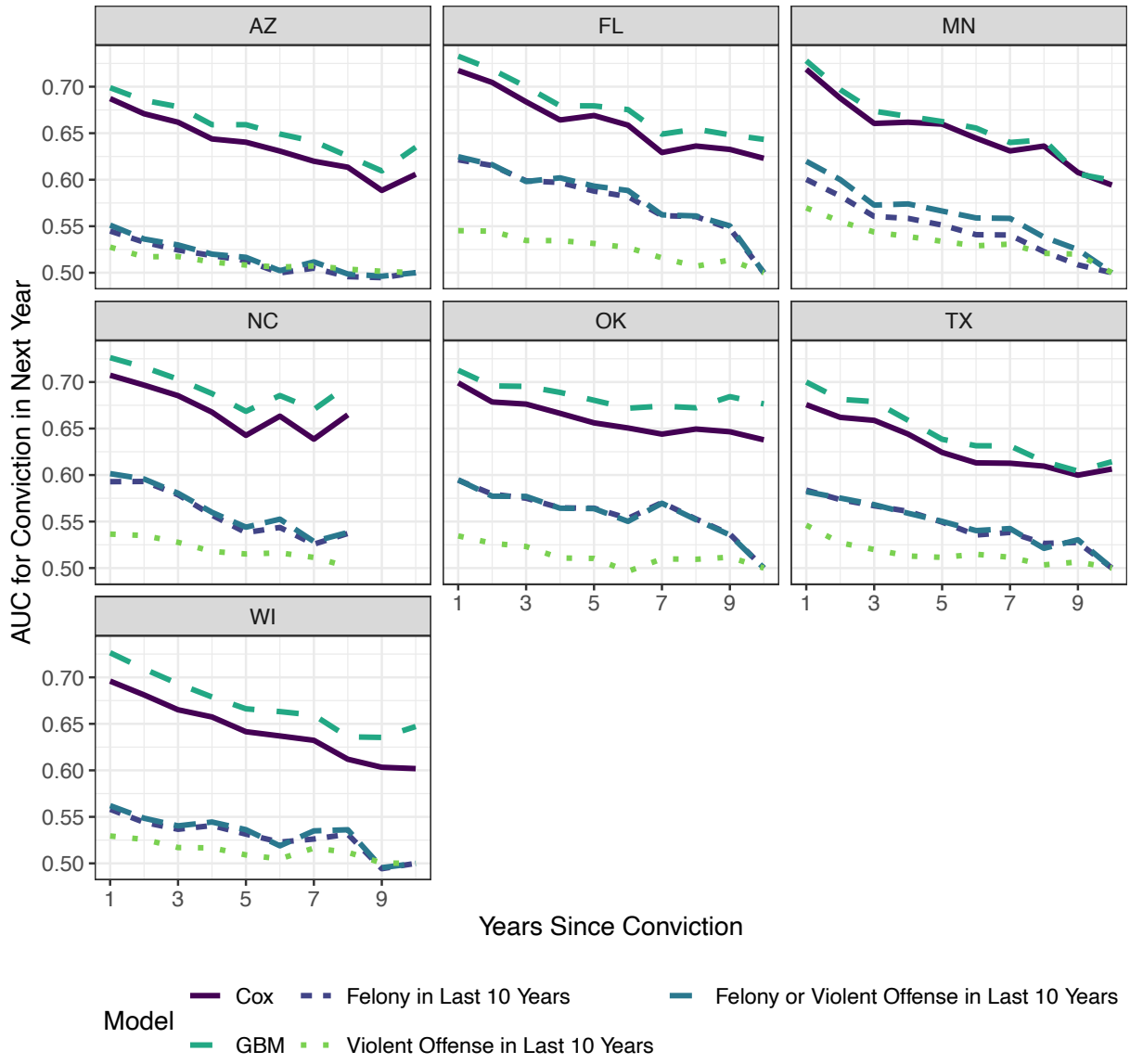


Figure S11: Area under the curve (AUC) for each model at a grid of years since conviction, stratified by state. Models under consideration include a Cox model, a generalized boosted model (GBM), and heuristics on whether the individual was convicted of a felony, violent offense, or either a felony or violent offense in the last 10 years. Higher values of the AUC are better.

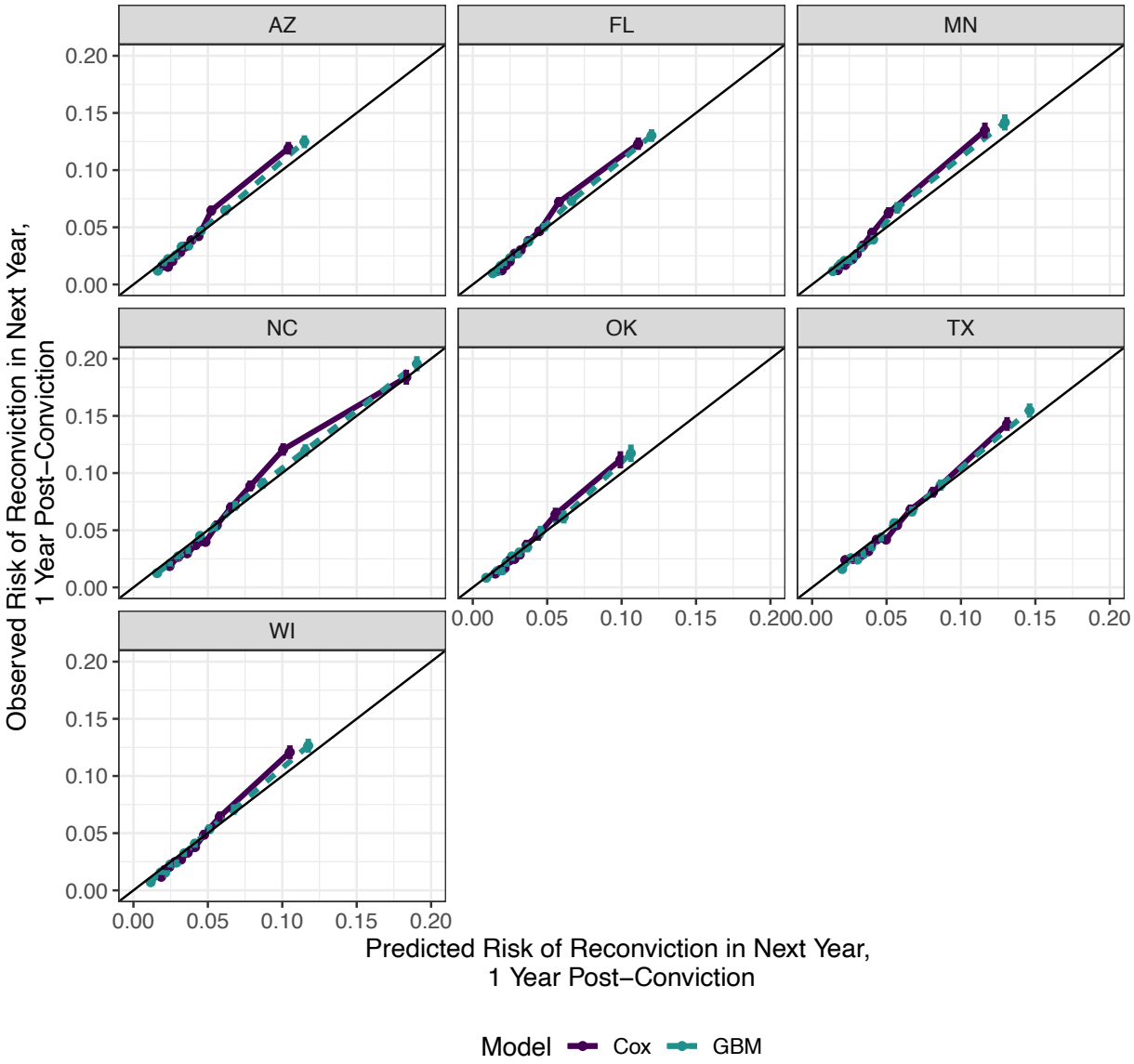


Figure S12: Calibration of a Cox proportional hazards model and generalized boosted model (GBM) for predicting recidivism one year after the most recent conviction, stratified by state. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated. The top left quadrant indicates that the model generated risk predictions that were lower than they should have been for the group (optimism); the bottom right quadrant indicates that risk predictions were higher than they should have been (pessimism).

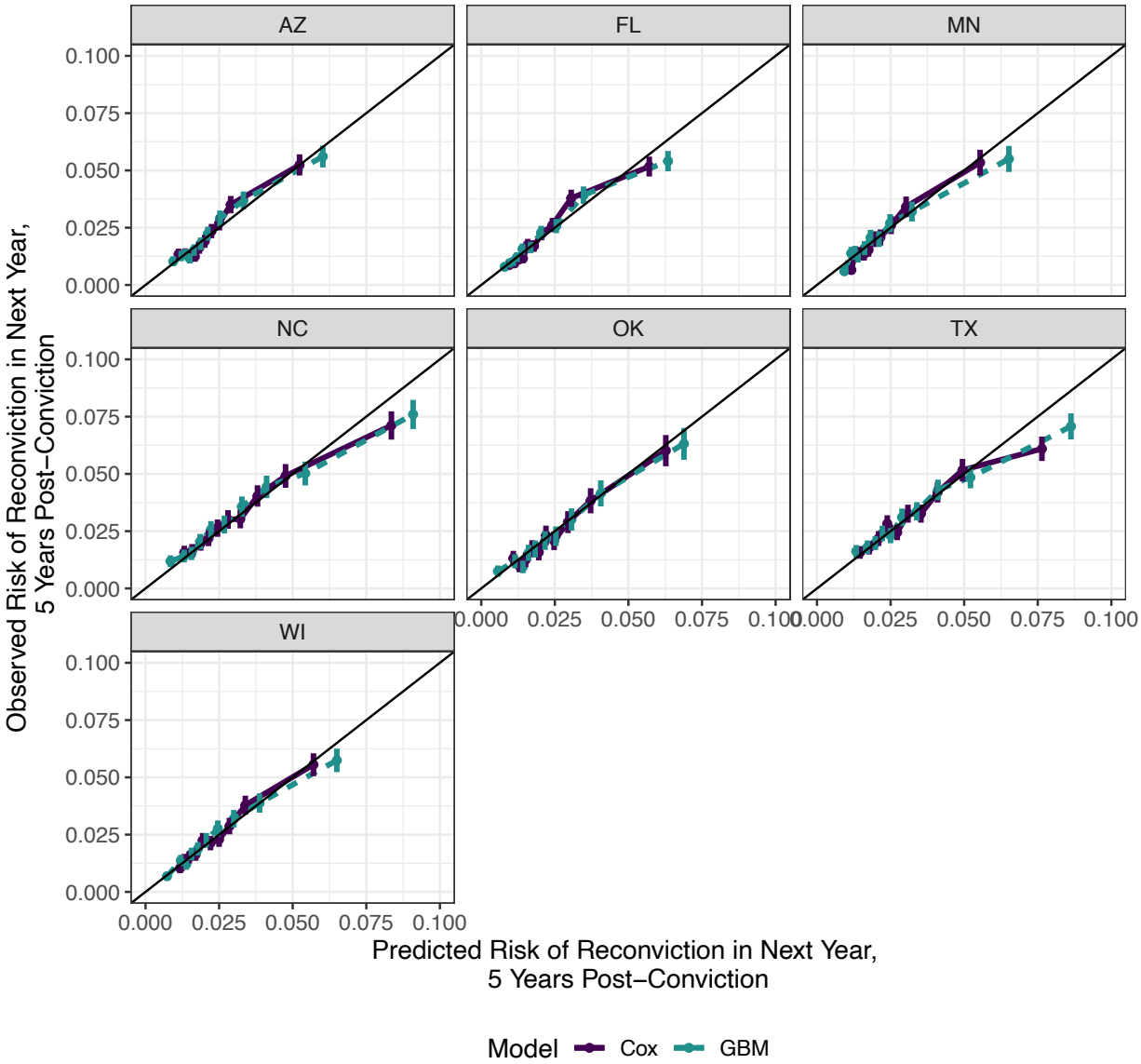


Figure S13: Calibration of a Cox proportional hazards model and generalized boosted model (GBM) for predicting recidivism five years after the most recent conviction, stratified by state. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated. The top left quadrant indicates that the model generated risk predictions that were lower than they should have been for the group (optimism); the bottom right quadrant indicates that risk predictions were higher than they should have been (pessimism).

4 Appendix 4: Additional Results on Algorithmic Fairness

Here we present some additional results on algorithmic fairness that could not be presented in the main manuscript due to space constraints.

Figures S14-S16 show the AUC of each of our candidate models, stratified by race and sex, to evaluate how predictive performance varies for different race and sex groups. Generally, there was more variability in predictive performance across timepoints for women than for men, particularly Hispanic women (who had relatively small sample sizes in the CJARS data). However, AUC was fairly comparable across race, sex, and state for all models under consideration.

Figure S17 and S18 compare the AUC and calibration for the Cox model in North Carolina, fit both with and without information on race and sex. Under federal guidance, a model containing race and sex could not be used for hiring decisions; these plots explore how much predictive performance is harmed by not including race and sex information in the model. For AUC, including race and sex only slightly affects the AUC, with race and sex providing improved predictive performance for Hispanic men and women and non-Hispanic white women. However, it has little effect for non-Hispanic Black men and women and non-Hispanic White women. For calibration, we see a similar but slightly different story, in which including race and sex information largely has no effect, but does offer slightly improved calibration for Hispanic men and women and non-Hispanic Black men.

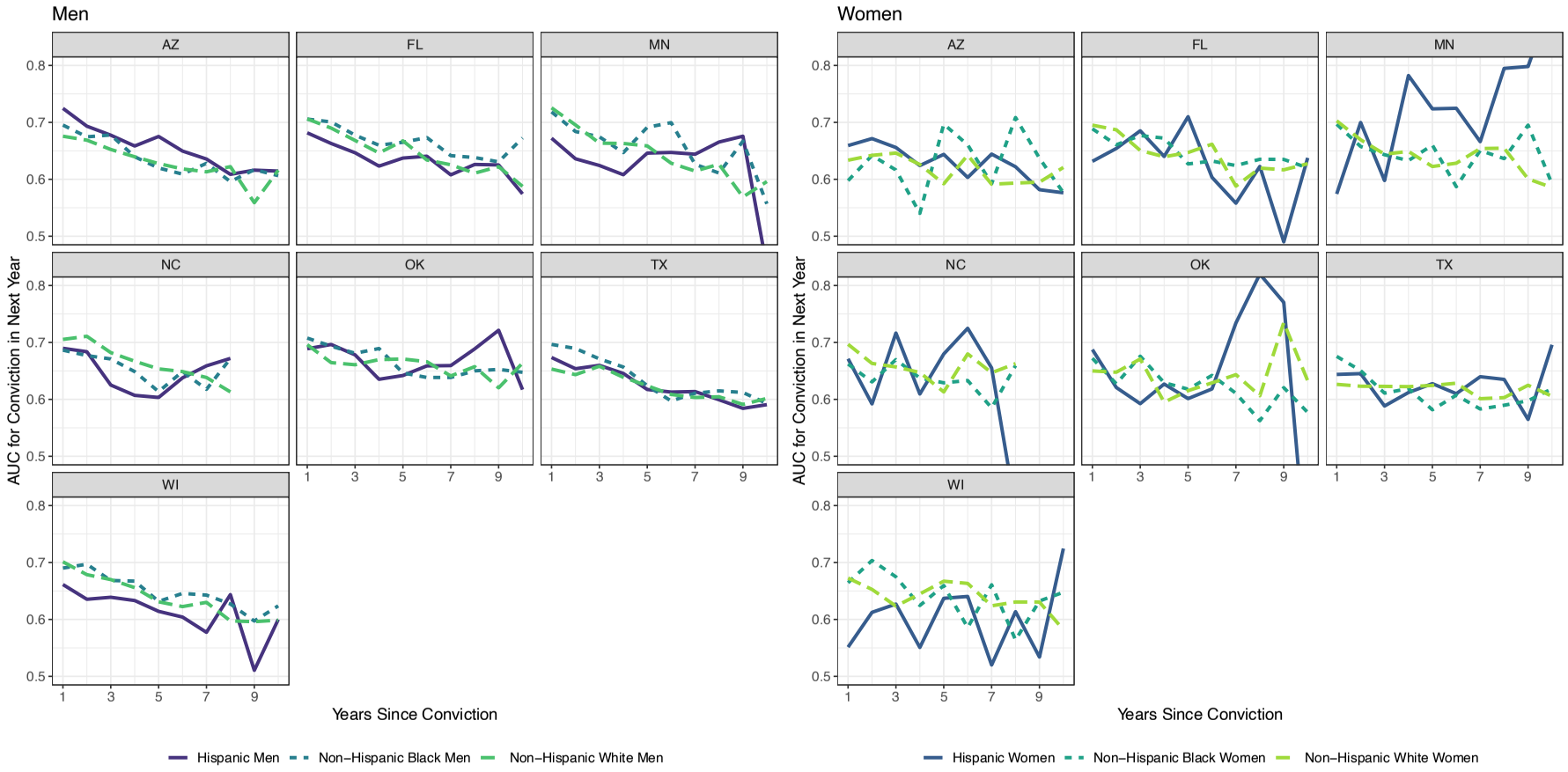


Figure S14: Area under the curve (AUC) for a Cox proportional hazards model at a grid of years since conviction, stratified by sex, state, and race. Higher values of the AUC are better. A perfectly fair model with respect to AUC would produce equivalent AUC for all race and sex groups at each timepoint.

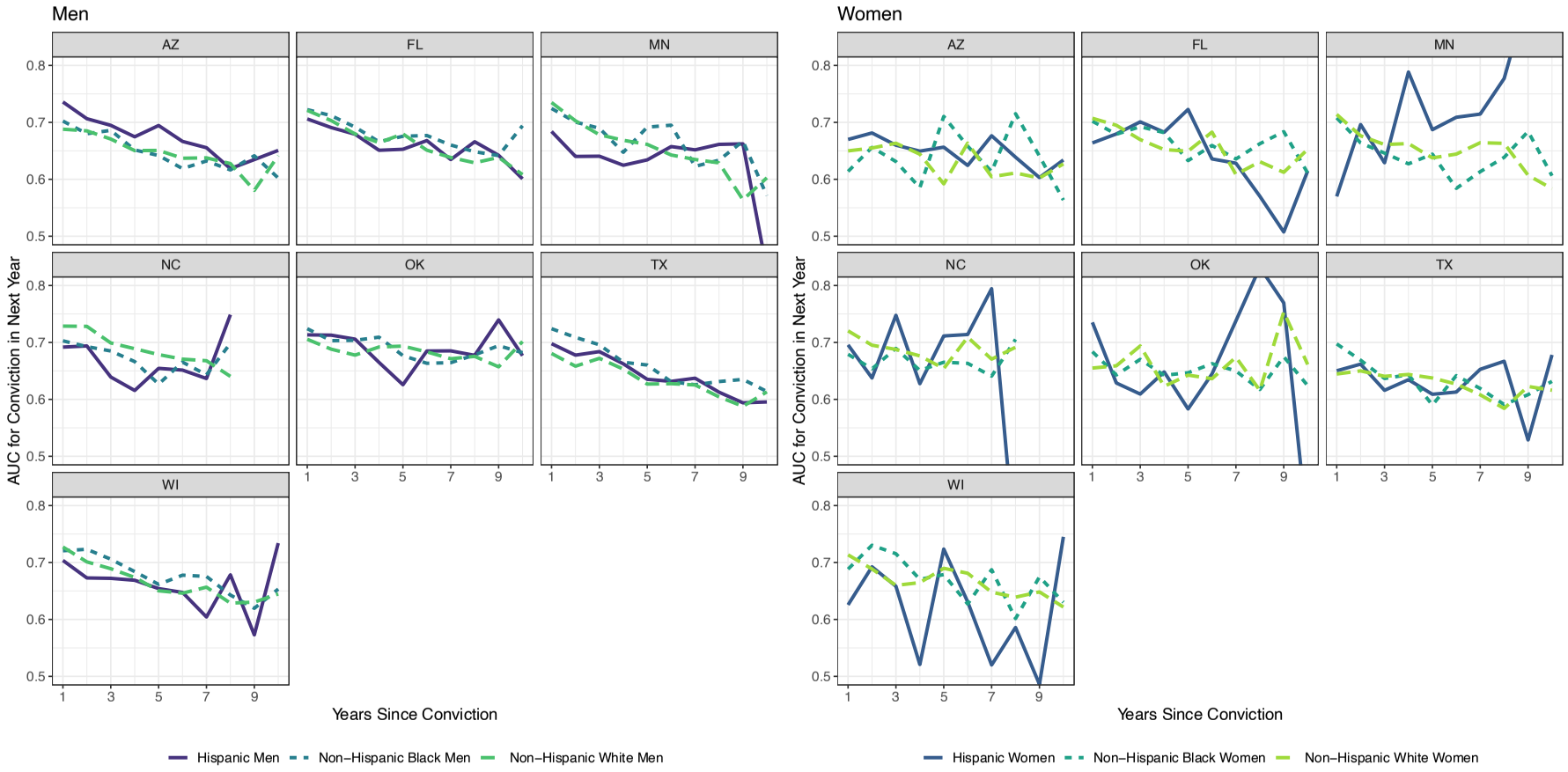


Figure S15: Area under the curve (AUC) for a generalized boosted model (GBM) at a grid of years since conviction, stratified by sex, state, and race. Higher values of the AUC are better. A perfectly fair model with respect to AUC would produce equivalent AUC for all race and sex groups at each timepoint.

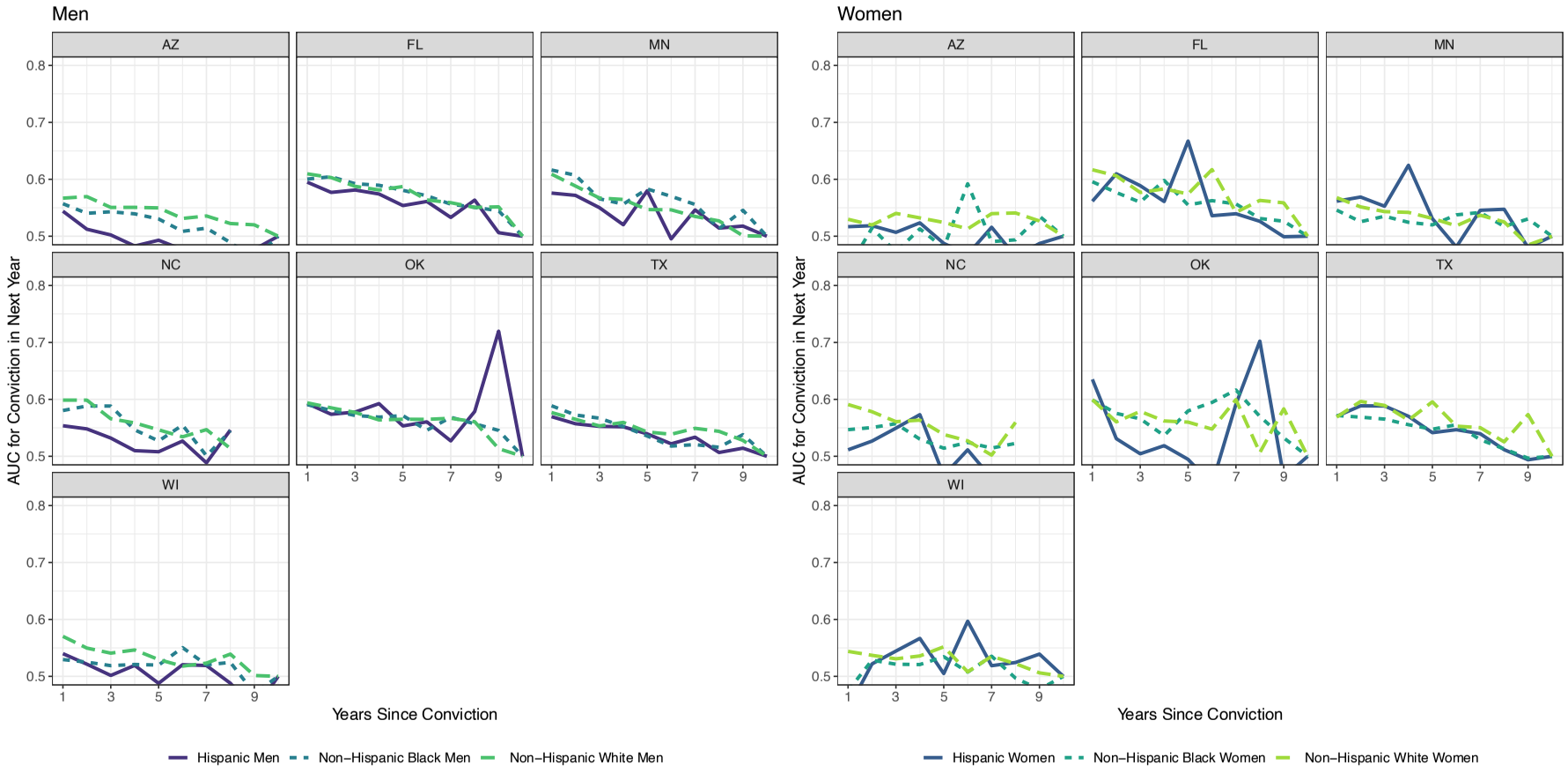


Figure S16: Area under the curve (AUC) for a heuristic of whether the person was convicted of a felony in the last 10 years, calculated at a grid of years since conviction, stratified by sex, state, and race. Higher values of the AUC are better. A perfectly fair model with respect to AUC would produce equivalent AUC for all race and sex groups at each timepoint.

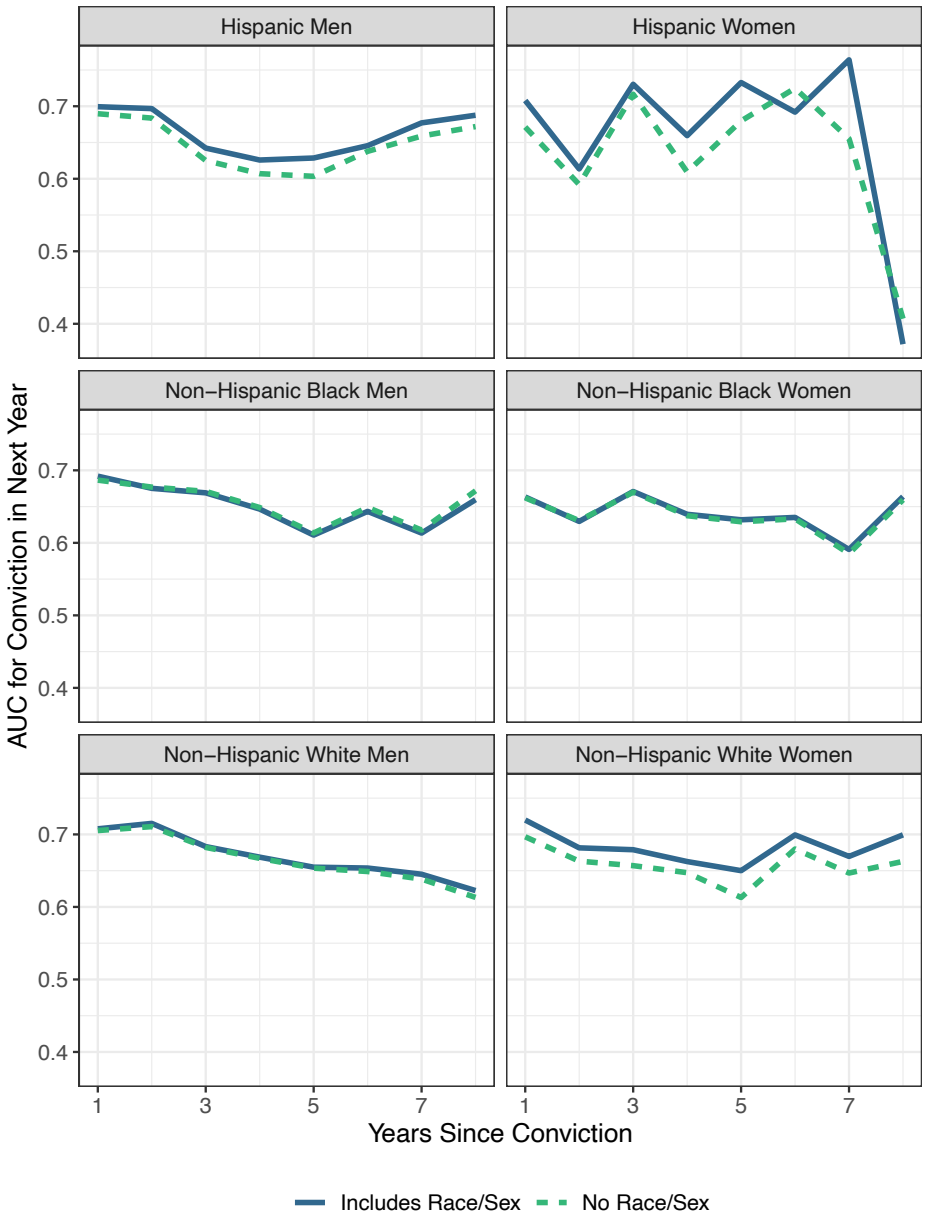


Figure S17: Area under the curve (AUC) for a Cox proportional hazards model fit in North Carolina data that adjusts for race and sex, and a second Cox model fit in North Carolina data that does not adjust for race and sex, calculated at a grid of years since conviction, stratified by sex and race. Higher values of the AUC are better. A perfectly fair model with respect to AUC would produce equivalent AUC for all race and sex groups at each timepoint.

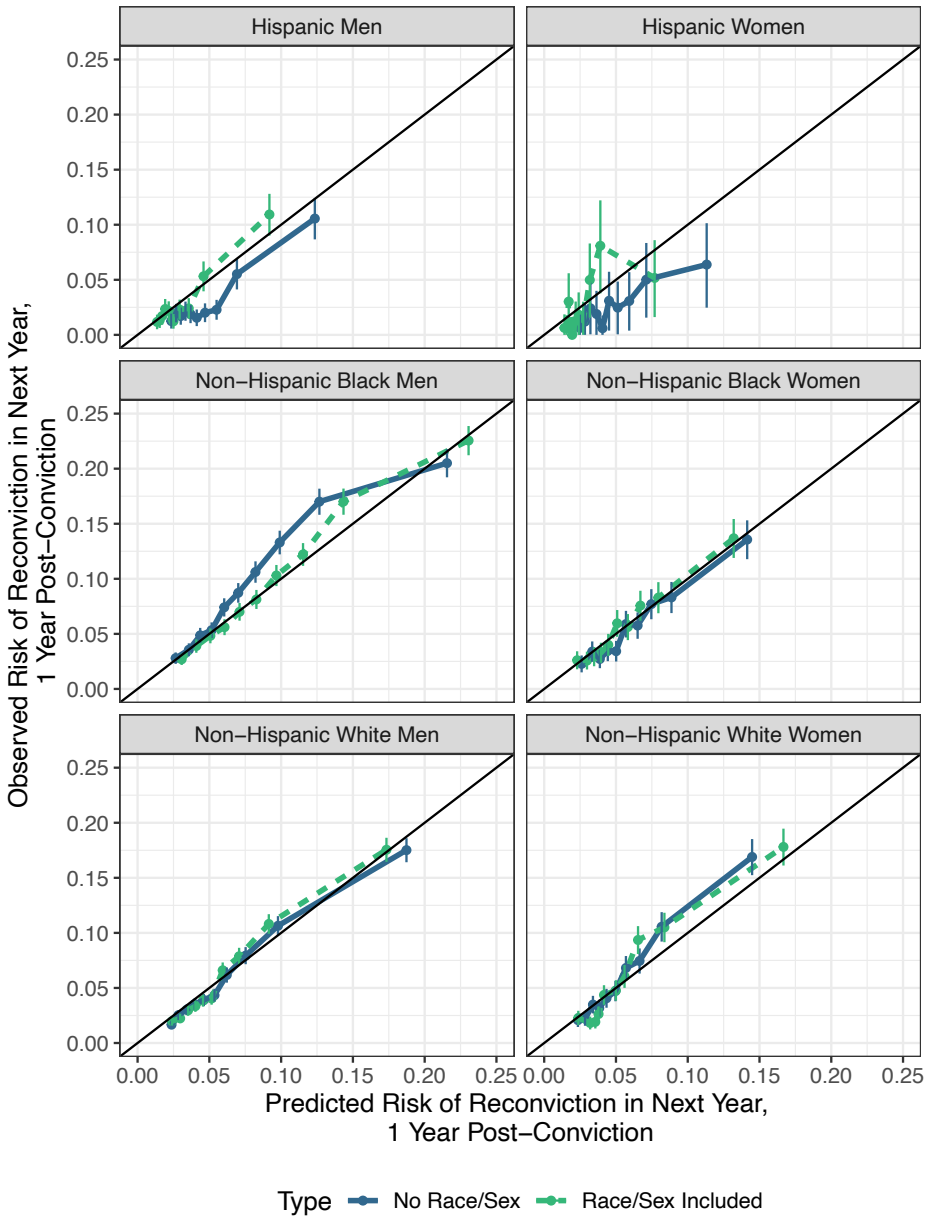


Figure S18: Calibration of a Cox proportional hazards model fit in North Carolina data that adjusts for race and sex, and a second Cox model fit in North Carolina data that does not adjust for race and sex, stratified by state, sex, and race. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated. A perfectly fair model with respect to calibration would produce equivalent calibration for all race and sex groups.

5 Appendix 5: Results on Model Portability

Here we present results on model portability across states. These figures were discussed in the main manuscript but are presented here due to space constraints.

Figure S19 shows the AUC in each state for 5 different Cox models: the Cox model fit using only Arizona data, a Cox model fit using equal samples from all 7 states (jumbo model), the Cox model fit only using Minnesota data, the Cox model fit using only North Carolina data, and the Cox model fit using the state's own data (own model). For all states, we unsurprisingly note that the model fit using the state's own data offers the best AUC. However, for many of the states (Florida, Minnesota, North Carolina, Oklahoma, Texas), the jumbo model offered decent predictive performance that was only slightly worse than the state's own model, suggesting that a jumbo model may be a good option. However, for Arizona and Wisconsin there were more substantial gaps between the state's own model and the jumbo model—in both cases, the Arizona model was a better fit than the jumbo model, suggesting that something may be distinct about Arizona and Wisconsin relative to the other 5 states.

Figure S20 and S21 present similar results, but for calibration at 1 and 5 years after the most recent conviction. For calibration, it is more difficult to transfer models across states, and the state's own model offers the best predictive performance. However, the jumbo model does provide adequate calibration for Arizona, Florida, North Carolina, Oklahoma, and Wisconsin. For Minnesota and Texas, the jumbo model generates predictions that seem consistently more optimistic than they should be (e.g. predicted risk is lower than observed risk). These issues attenuate somewhat at the 5 year mark after the most recent conviction.

Figure S22 shows how the jumbo model compares relative to the state's own model in terms of ranked risk. Although ranks are changed by the jumbo model, these changes are generally minor.

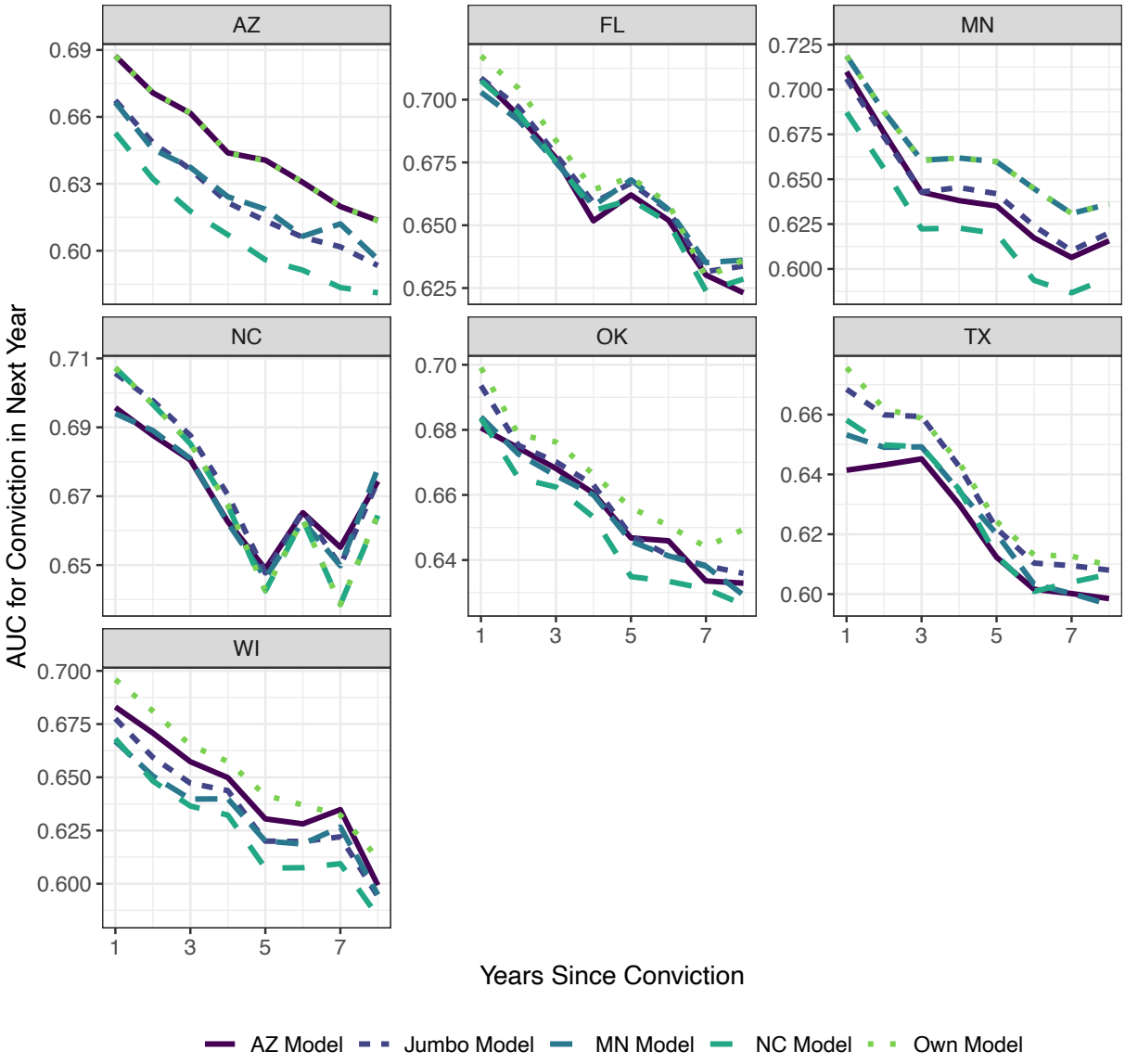


Figure S19: Area under the curve (AUC) for each model at a grid of years since conviction, stratified by state. Models under consideration include a Cox model fit in Arizona data, a Cox model fit using equal samples of data from all states (jumbo model), a Cox model fit in Minnesota data, a Cox model fit in North Carolina data, and a Cox model fit using the state's own data. Higher values of the AUC are better.

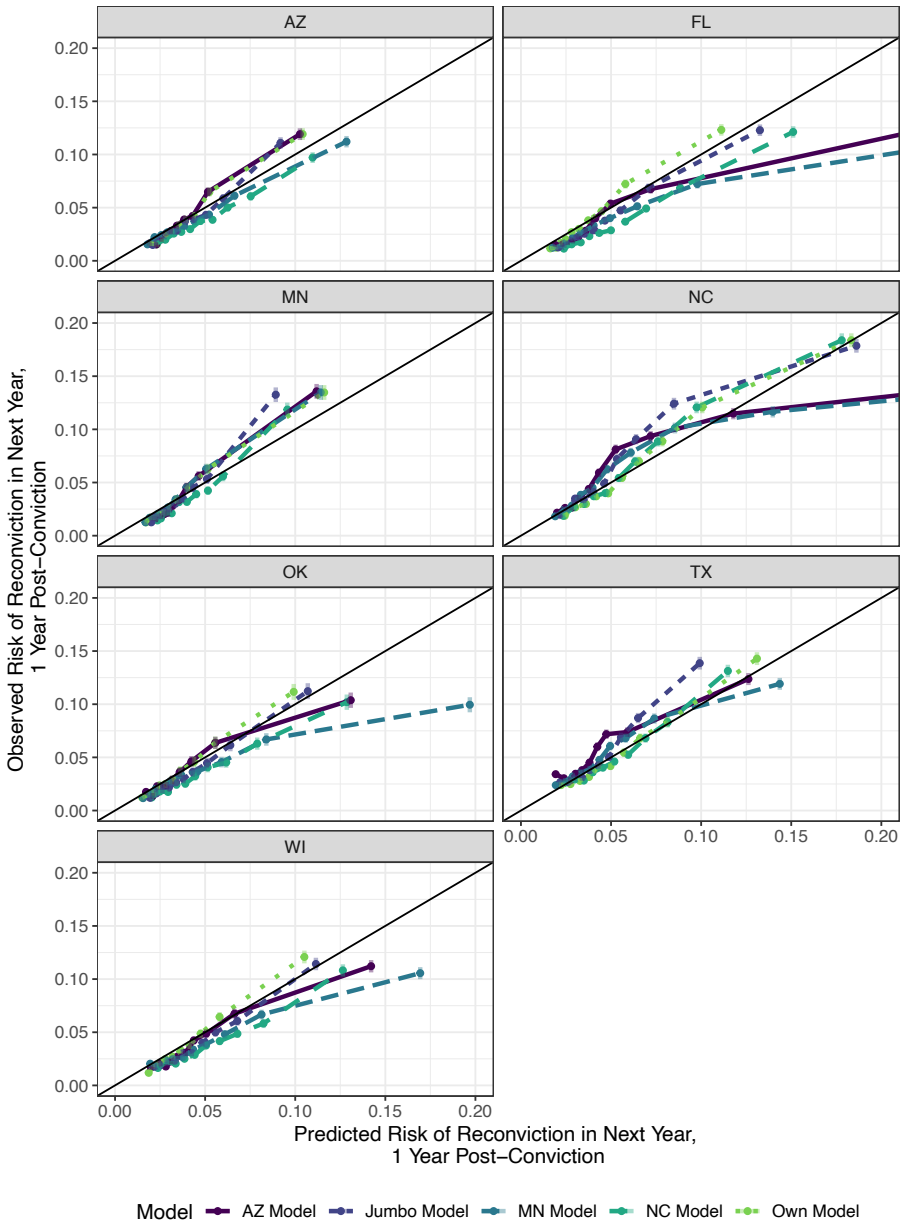


Figure S20: Calibration of several models for predicting recidivism one year after the most recent conviction, stratified by state. Models under consideration include a Cox model fit in Arizona data, a Cox model fit using equal samples of data from all states (jumbo model), a Cox model fit in Minnesota data, a Cox model fit in North Carolina data, and a Cox model fit using the state's own data. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated.

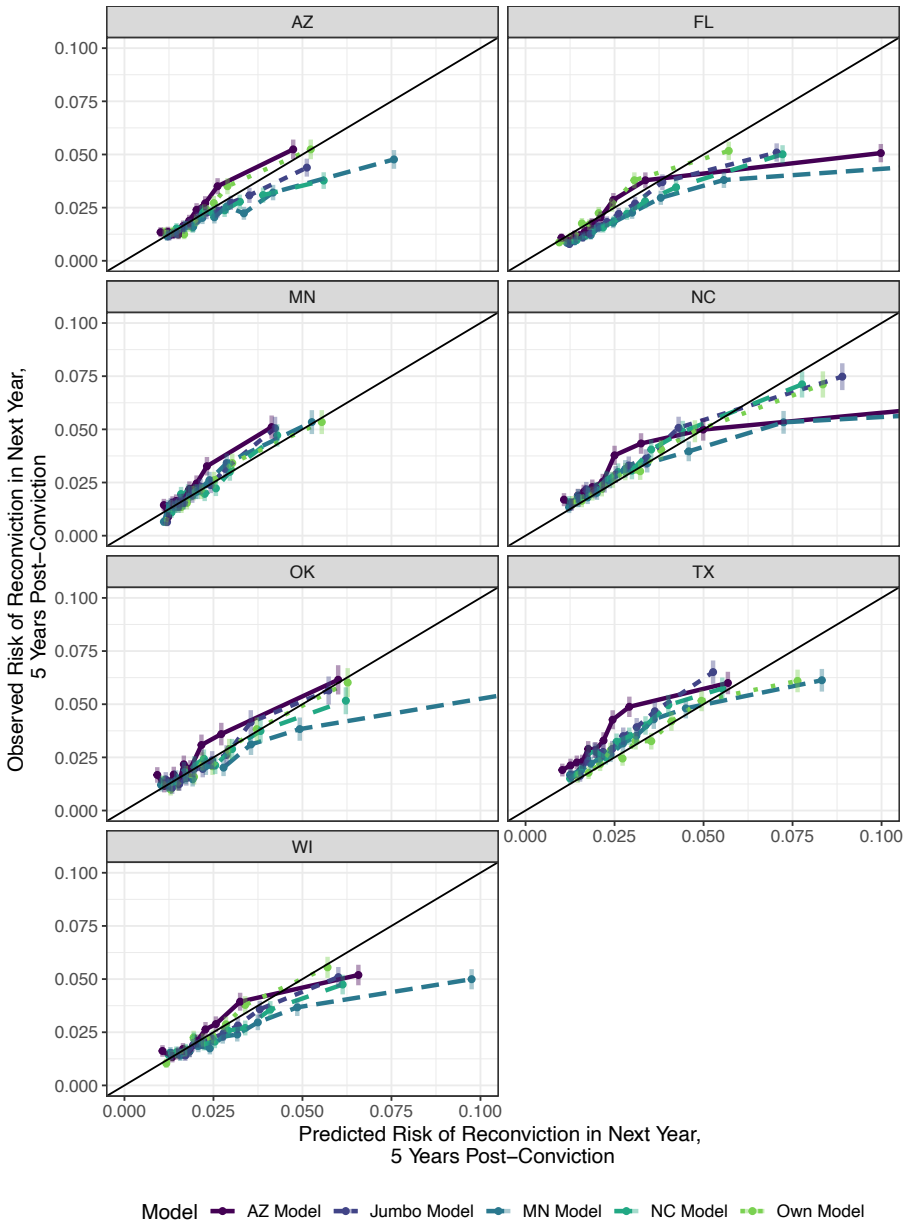


Figure S21: Calibration of several models for predicting recidivism 5 years after the most recent conviction, stratified by state. Models under consideration include a Cox model fit in Arizona data, a Cox model fit using equal samples of data from all states (jumbo model), a Cox model fit in Minnesota data, a Cox model fit in North Carolina data, and a Cox model fit using the state's own data. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated.

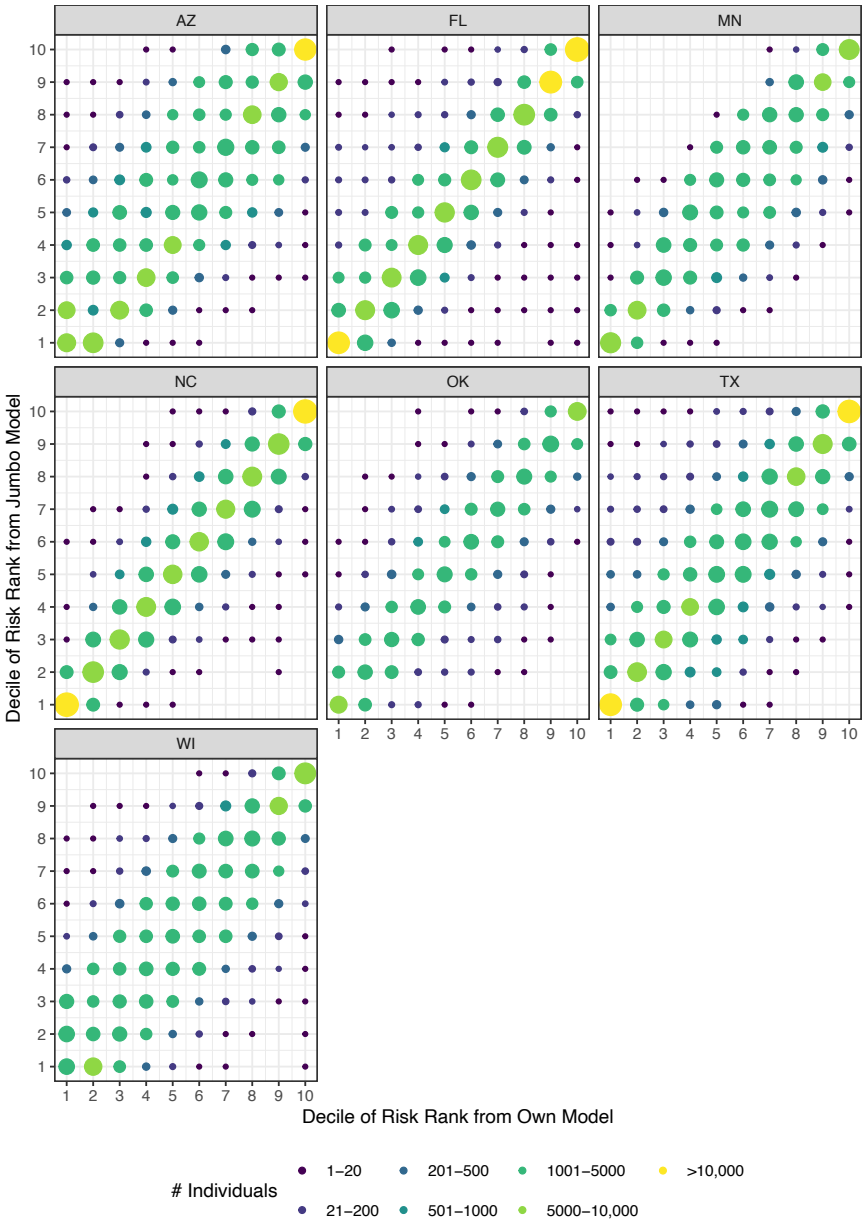


Figure S22: Comparison of ranked risk between each state's own Cox model and a Cox model fit using equal samples of data from all states (jumbo model). Each individual's risk was calculated from the two models, then ranked and split into deciles of ranked risk. Ideally all individuals would fall along the 45 degree diagonal in which risk rank decile does not change between models.

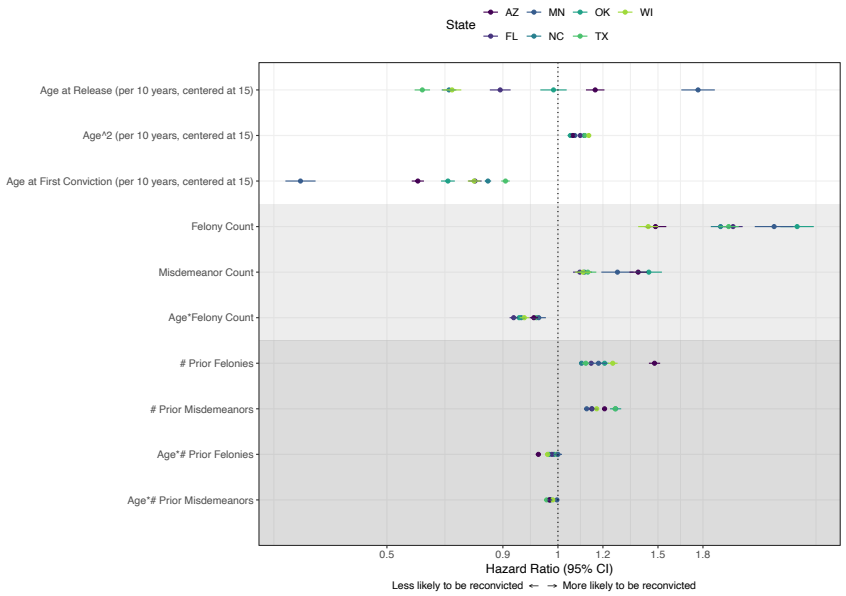


Figure S23: Hazard ratio and 95% confidence interval estimates for predicting recidivism using a Cox proportional hazards model fit using a reduced set of candidate predictors in a training sample of criminal records from the Criminal Justice Administrative Records System data (CJARS), stratified by state. Data range from 1992-2021, depending on the state.

6 Appendix 6: Results on Reduced Model

Finally, we present some results on a reduced Cox model that does not consider incarceration or crime type information. We fit this reduced model because many employers will not have access to this information as part of a routine criminal background check, and we wanted to explore how model performance is affected.

Figure S23 shows the coefficient estimates for the reduced Cox model. They are largely unchanged from the coefficient estimates presented for the full model in the main manuscript (Figure 2). Figure S24 compares the AUC of the full Cox model to the AUC of the reduced model. AUC does decrease when incarceration and crime type information are excluded from the model. Figure S25 compares the calibration of the two models. Finally, Figure S26 shows the calibration of the reduced model, stratified by race and sex, to see how removing incarceration/crime type information affects algorithmic fairness.

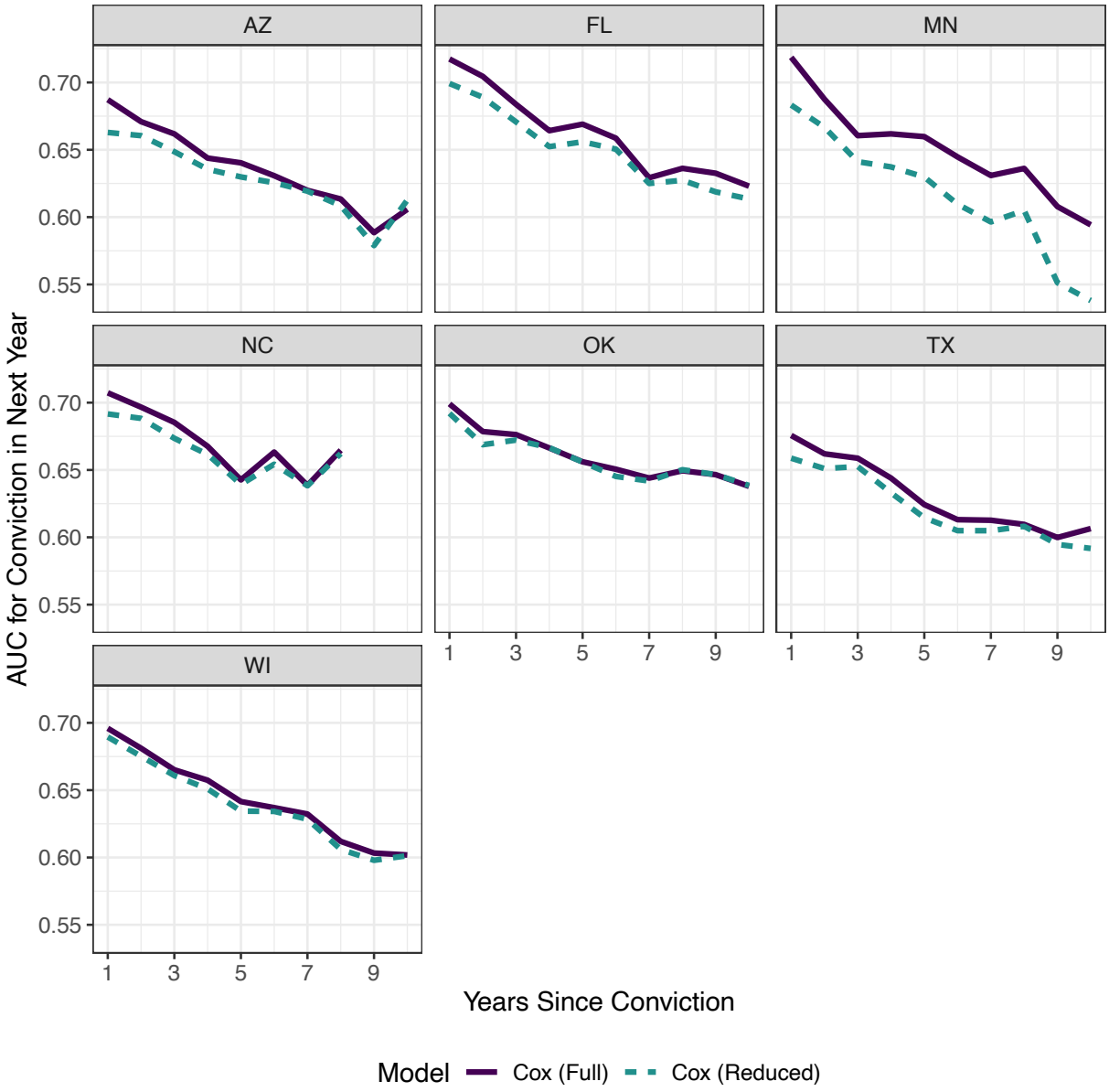


Figure S24: Area under the curve (AUC) for each model at a grid of years since conviction, stratified by state. Models under consideration include a Cox model using the full predictor set and a Cox model using a reduced set of predictors (no incarceration or crime type information). Higher values of the AUC are better.

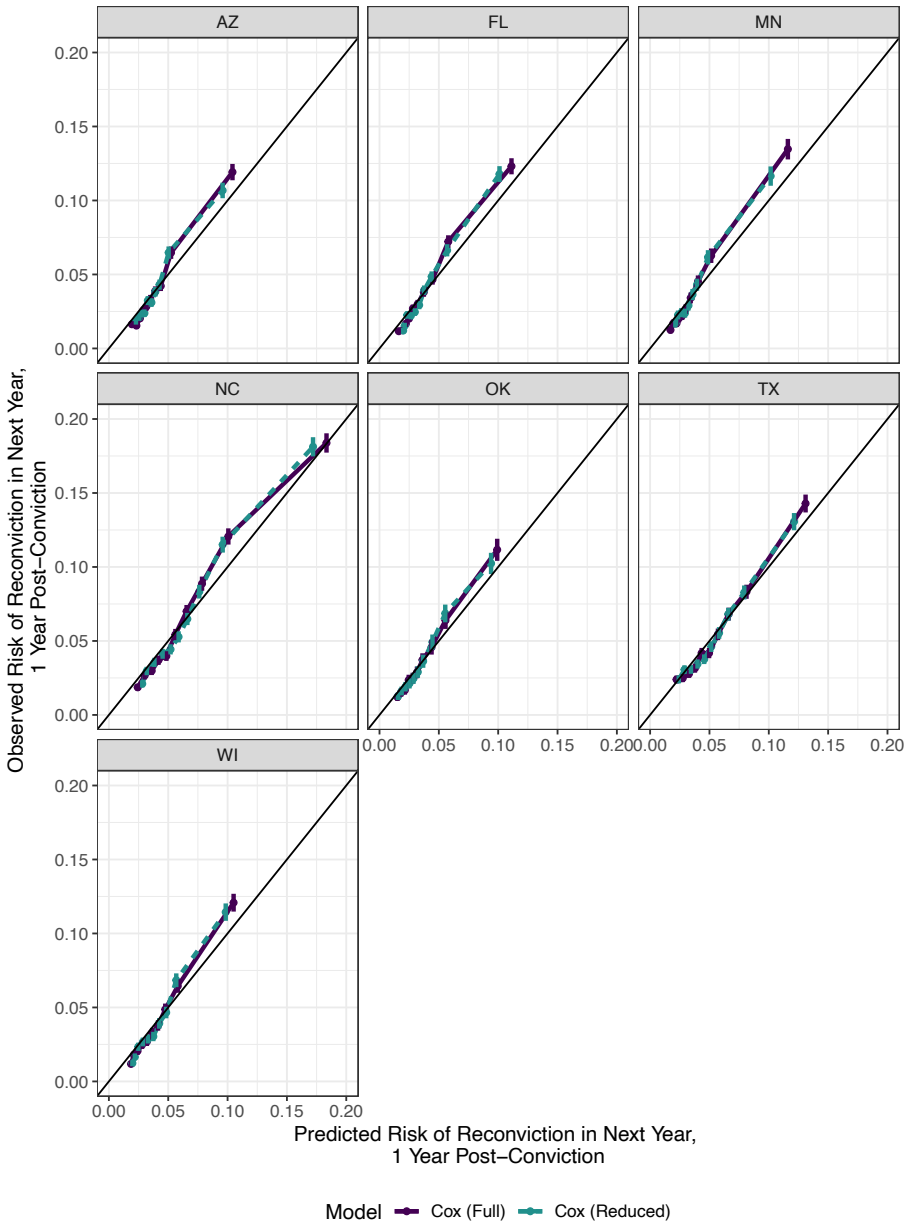


Figure S25: Calibration of a Cox proportional hazards model fit using the full set of predictors and a Cox proportional hazards model fit using a reduced set of predictors (no incarceration or crime type information) for predicting recidivism one year after the most recent conviction, stratified by state. Perfect calibration occurs along the 45 degree black line; a model with calibration further from the 45 degree line is worse calibrated. The top left quadrant indicates that the model generated risk predictions that were lower than they should have been for the group (optimism); the bottom right quadrant indicates that risk predictions were higher than they should have been (pessimism).

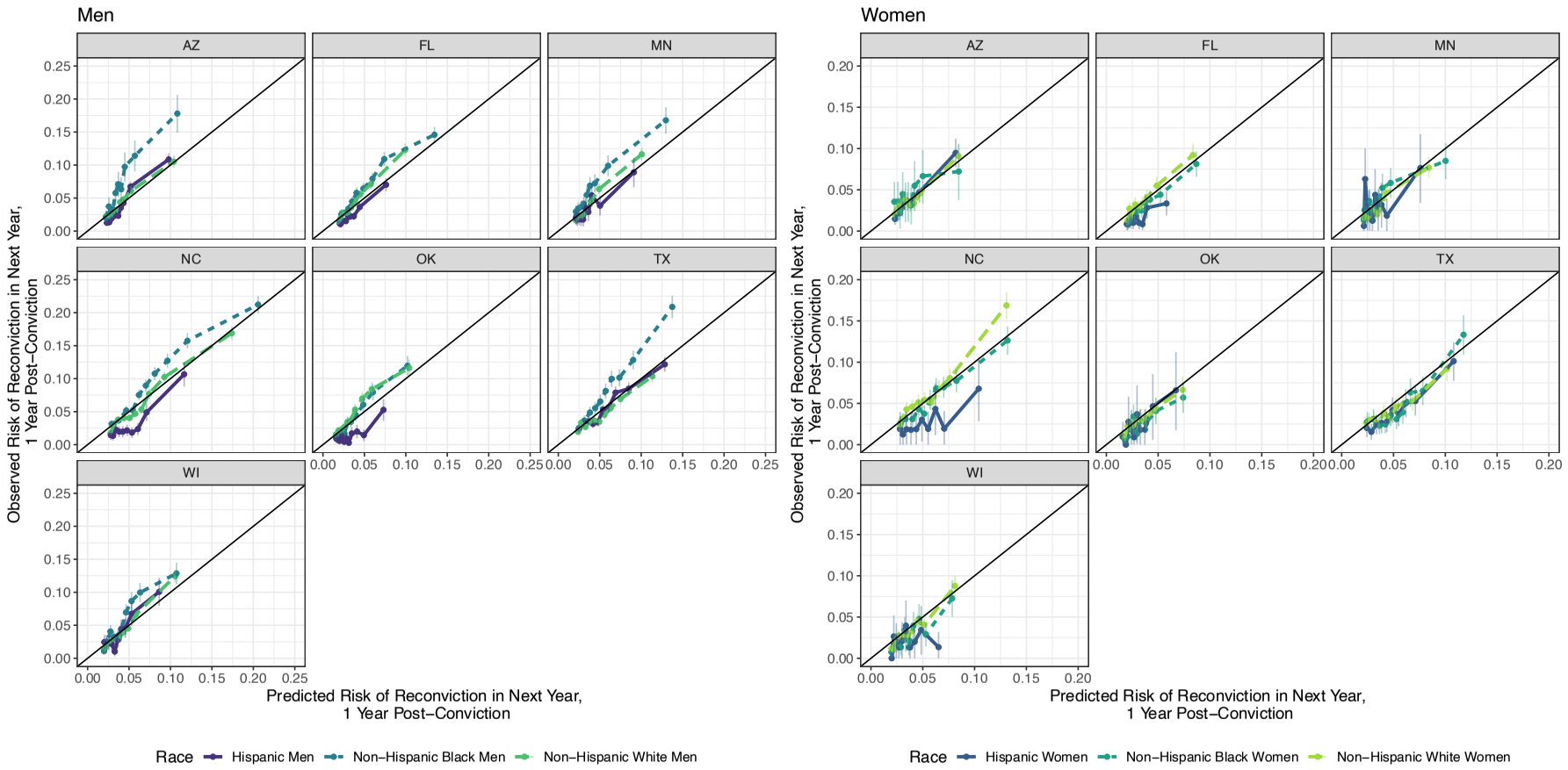


Figure S26: Race- and sex-specific calibration of a Cox proportional hazards model for predicting recidivism using a reduced set of predictors (no incarceration or crime type information), stratified by state. Perfect calibration occurs along the 45 degree black line; groups with calibration further from the 45 degree line are worse calibrated. The top left quadrant indicates that the model generated risk predictions that were lower than they should have been for the group (optimism); the bottom right quadrant indicates that risk predictions were higher than they should have been (pessimism).

References

SD Bushway, BG Vegetabile, N Kalra, L Remi, and G Baumann. Providing another chance: Resetting recidivism risk in criminal background checks. Research Report RRA1360-1, RAND, 2022.