



Scaling Pretrial Risk Assessment: A Statewide Validation of the Public Safety Assessment (PSA) Across New Mexico's Judicial Districts

Prepared By:

Alex Severson, Ph.D.

Elise Ferguson, M.A.

Daniel Goldberg, M.A.

Prepared For:

The New Mexico Administrative Office of the
Courts (AOC)

December 2025

Executive Summary

Background and Purpose

The Public Safety Assessment (PSA) is an evidence-based risk assessment instrument (RAI) designed to provide judicial decision-makers with information that may help inform pretrial release and conditions of supervision decisions. Among other things, the **PSA uses** defendant characteristics and criminal history information to generate predictions across three domains: the likelihood of failure to appear in court, the likelihood of engaging in new criminal activity during the pretrial period, and the likelihood of committing violent offenses while awaiting trial. New Mexico first implemented the PSA in Bernalillo County in 2017 and has, through December 2025, expanded its use to 11 additional judicial districts throughout the state. The current study examines the PSA's predictive performance (i.e., how well scores on the tool correlate with pretrial success, defined as showing up for court appearances and not engaging in new criminal activity upon release before case disposition) across New Mexico judicial districts that have implemented the PSA.

Key Findings

Specifically, we analyzed 10,872 felony cases from two judicial districts (i.e., the First and Thirteenth Judicial Districts), which had sufficiently reliable court and jail data. Our analysis indicated that the PSA exhibited moderate to good performance for predicting court appearance, with successful appearance rates declining from 90% among defendants classified as lowest risk for failure to appear to 48.5% for those classified as highest risk. Similar patterns emerged for predicting new criminal activity, with public safety rates declining from 92% in the lowest-risk category to approximately 60% in the highest-risk category. However, the PSA had lower accuracy in predicting violent criminal activity, a result consistent [with findings from our 2024 revalidation study of the PSA in the Second Judicial District](#). Regression models confirmed that the associations between PSA risk scores and all three outcome measures remained even after accounting for defendant demographics, supervision intensity, exposure duration, and court-jurisdiction clustering effects.

Limitations

It is important to be clear about the limitations of the current study. Race and ethnicity data were unavailable for 64.6% of the sample, so we could not evaluate differences in the predictive performance of PSA across racial and ethnic groups. Secondly, data quality issues in jail booking and release records meant that we had to limit our primary analysis to two of six PSA-implementing districts, reducing the sample from 53,268 potential cases to 10,872, and potentially constraining how exportable our findings are across the remaining PSA-implementing jurisdictions, which may differ in systematic ways from the First and Thirteenth Judicial Districts.

Recommendations

We recommend continued use of the PSA as a decision-support instrument for pretrial release and supervision determinations in New Mexico, coupled with efforts to address identified data infrastructure limitations in other implementing jurisdictions. Specifically, jails and courts should implement standardized protocols for collecting and recording booking information and demographic data to enable more comprehensive validation analyses. Regular revalidation studies would facilitate ongoing monitoring of the PSA's performance as New Mexico's pretrial population and criminal justice practices evolve. Additionally, decision-makers should receive more information regarding prediction uncertainty, particularly for violent crime outcomes where the PSA's discriminative ability was more constrained. The PSA should be understood and applied as one component of a comprehensive decision-making framework that integrates actuarial information with judicial discretion and case-specific considerations, rather than as a deterministic prediction tool that substitutes for professional, case-by-case judgment or for the consideration of contextual factors not captured in the PSA.

Introduction

The Public Safety Assessment (PSA) is an evidence-based judicial decision-making tool developed by Arnold Ventures designed to assist judges in making release decisions during a defendant's pretrial phase (AdvancingPretrial.org, 2020). Arnold Ventures originally constructed the PSA using data from approximately 750,000 cases from 300 jurisdictions to develop a scored set of risk factors predictive of an individual's likelihood of appearing in court (i.e., the FTA or Failure to Appear rate) and engaging in new criminal activity during the pretrial period (i.e., the NCA or New Criminal Activity rate). The PSA also flags individuals who present an elevated risk of committing a violent crime during the pretrial period (i.e., the NVCA flag). The PSA has been validated on over half a million cases nationally and has been revalidated in diverse locations, including Kentucky (DeMichele et al., 2018), California (Skog & Lacoce, 2021), and Texas (Greiner et al., 2020).

New Mexico's implementation of the PSA began in Bernalillo County in June 2017, where it was implemented exclusively for felony cases, making it unique among other jurisdictions that typically use the PSA to evaluate risk for both misdemeanants and felons. Following the initial implementation and subsequent validation studies demonstrating the tool's effectiveness in Bernalillo County (Ferguson et al., 2021; [Severson & Ferguson, 2024](#)), the New Mexico Administrative Office of the Courts (AOC) began expanding PSA implementation to other judicial districts throughout the state, beginning in approximately February 2020. This statewide expansion represents a significant scaling effort that provides a unique opportunity to examine how the PSA performs across different jurisdictional contexts within a unified court system. Through the study period timeline (i.e., June 2023), the PSA had been implemented in six additional judicial districts beyond the 2nd Judicial District (Bernalillo County), encompassing the 1st Judicial District (Santa Fe, Rio Arriba, and Los Alamos counties), 3rd Judicial District (Doña Ana County), 4th Judicial District (San Miguel and Mora counties), 6th Judicial District (Luna, Grant, and Hidalgo counties), 11th Judicial District (San Juan and McKinley counties), and 13th Judicial District (Sandoval, Cibola, and Valencia counties). As of December 2025, the PSA has been implemented in 12 of 13 judicial districts.

The expansion of the PSA across a majority of New Mexico's judicial districts presents both opportunities and challenges for evaluating the predictive validity of the PSA. While the PSA was designed to provide consistent risk assessment across jurisdictions, local validation is still essential to ensure the PSA's predictive validity and fairness within specific populations and contexts (National Association of Pretrial Services Agencies, 2020). Each judicial district in New Mexico serves distinct geographic regions with varying demographic compositions, case volumes, court procedures, resource availability, and variable release decision-matrices (RDMs), factors that may influence both the implementation of the PSA and the predictive performance of the PSA. Moreover, the rural-urban divide that characterizes much of New Mexico's geography means that some districts serve predominantly urban populations, whereas others serve more rural and tribal communities, which may jointly influence baseline outcome rates and the observed relationship between PSA risk factors and pretrial outcomes.

In what follows, we briefly review the existing research on the PSA's predictive validity to contextualize our findings within the broader evidence landscape. We then discuss our sampling criteria and data collection procedures, followed by a review of our key results. Specifically, we first provide an overview of descriptive statistics for the full statewide sample (i.e., all included implementing jurisdictions), including demographic characteristics and case composition across all participating districts. Next, we present an overview of failure to appear (FTA), new criminal activity (NCA), and New Violent Criminal Activity (NVCA) PSA score distributions by judicial district and individual courts to examine baseline patterns in risk assessment scoring. We then present results of predictive validity tests which evaluate the association between FTA, NCA, and NVCA scores with observed and predicted Court Appearance Rates (CAR), Public Safety Rates (PSR), and Non-Violence Rates (NVR), using area under the curve (AUC) analyses and logistic random-effects regression models, focusing on the two districts, specifically, the 1st and 12th judicial districts, which had the highest-quality reported jail booking release log and court system data. Following this, we evaluate variation in predictive performance by sex across districts to assess potential bias in the PSA's predictive performance across demographic groups. Finally, we conclude with a discussion of limitations to the current study, implications for policy and practice, and directions for future research.

Literature Review

Pretrial Risk Assessment and Cross-Jurisdictional Validity

The development and implementation of standardized pretrial risk assessment instruments (PRAIs) has been a significant focus of criminal justice reform efforts over the past two decades. Research consistently demonstrates that PRAIs outperform clinical judgment in predicting criminal justice outcomes, like violent and sexual reoffending and pretrial criminal activity (Slobogin, 2023; Goossens et al., 2024). However, the accuracy of these tools varies across jurisdictions, populations, and implementation contexts. A growing body of literature emphasizes the importance of local validation studies to ensure that PRAIs maintain their predictive validity when applied to new populations or geographic contexts outside of the initial validation sample (Skeem & Lowenkamp, 2016; Fazel et al., 2022). This research suggests that tools developed using data from one jurisdiction may not perform equally well in another, even when the underlying risk factors measured appear theoretically sound.

Studies examining cross-jurisdictional performance of risk assessment tools have identified several factors that can influence estimates of predictive validity across different implementation contexts. Demographic composition represents one of the most significant factors, as risk assessment tools developed on samples with different racial, ethnic, or socioeconomic compositions may not generalize effectively to populations with different characteristics (Flores et al., 2016). Geographic factors also play a role, as rural and urban jurisdictions often exhibit different patterns of criminal behavior, court processing times, and available community resources that can affect both baseline failure rates and the predictive usefulness of specific risk factors (Bornstein et al., 2013). Court procedural differences, including variations in hearing schedules, case processing timeframes, and available pretrial services, can also influence the relationship between assessed risk and observed pretrial outcomes (Lowenkamp et al., 2013). Beyond court procedures, stakeholder practices outside the courtroom warrant consideration, as prosecutorial decision-making on charge filing, charging severity, and dismissal timing can shape outcomes that risk assessment tools aim to predict. For example, previous research indicates that prosecutorial discretion and specifics about charging scope vary substantially across jurisdictions based on local policies, resource constraints, and organizational norms, potentially affecting both baseline rates of pretrial success and, in consequence, the performance of standardized assessment instruments (National Institute of Justice, 2014; Harvard Law Review, 2023). Moreover, to the extent that prosecutorial choices have downstream impacts on time to case disposition, these can, in turn, affect exposure duration (i.e., the time in the community), which can, in turn, affect potential pretrial success rates. Together, this body of research highlights the importance of conducting jurisdiction-specific validation studies, particularly when implementing standardized tools across contexts that vary in several ways that may impact the association between PSA scores and outcomes.

Public Safety Assessment Validation Research

PSA Validation Research Outside of New Mexico

As noted earlier, the PSA has been subject to extensive validation research across multiple jurisdictions since its initial development, with the existing literature generally reporting that the PSA has reasonable levels of predictive validity for predicting the array of pretrial outcomes it is designed to predict (i.e., FTA; NCA; NVCA). The original validation study conducted by Arnold Ventures used data from over 750,000 cases across 300 jurisdictions, establishing baseline performance metrics and demonstrating the tool's general effectiveness across diverse contexts (Laura & John Arnold Foundation, 2016). Subsequent independent validation studies have examined the PSA's performance in specific jurisdictions, with results generally supporting the tool's predictive validity while also revealing important variation across different implementing contexts. For example, DeMichele et al. (2018) conducted a PSA validation study in Kentucky and found that the PSA demonstrated good predictive validity overall (AUC scores ranging from 0.67 to 0.70) while also identifying some differences in performance across racial groups that warranted further investigation.

California's validation experience has also provided valuable insights into the challenges of implementing the PSA across diverse jurisdictions. Skog and Lacoé (2021) examined PSA implementation in San Francisco and reported that while the PSA demonstrated adequate predictive validity, local factors such

as differences in charging practices and court procedures required modifications to implementation protocols. Similarly, validation studies in other jurisdictions have found that while the PSA's core structure remains valid, local adaptations and ongoing monitoring are essential for optimal performance (Greiner et al., 2020). These studies have also highlighted the importance of examining predictive fairness across different demographic groups, as some research has identified variations in predictive accuracy that may contribute to disparate outcomes of pretrial decision-making (Angelino et al., 2017). In sum, the available evidence through 2025 suggests that while the PSA provides a comparatively robust foundation for pretrial risk assessment, successful implementation requires careful attention to local context and ongoing validation efforts.

PSA Validation Research in New Mexico

New Mexico's experience with the PSA began with Bernalillo County's implementation in 2017, followed by a comprehensive validation study completed by the University of New Mexico's Center for Applied Research and Analysis (CARA) in 2021. Ferguson et al. (2021) examined 10,289 felony PSA cases spanning June 2017 to March 2020, and reported that the PSA demonstrated fair to good predictive validity in Bernalillo County with AUC scores of 0.64 for both FTA and NCA outcomes and 0.57 for NVCA outcomes. Ferguson et al., (2021) also examined predictive fairness across racial and gender groups, finding no statistically significant differences in predictive validity by race and comparable performance across gender groups, suggesting reasonable calibration of the tool for the 2nd Judicial District's pretrial population.

A subsequent revalidation study in Bernalillo County using updated court data through June 2023 replicated the finding of the general predictive validity of the PSA in Bernalillo County while also identifying some longitudinal changes in failure rates and differences in the association between different PSA risk factors with pretrial outcomes of interest (Severson & Ferguson, 2024). This PSA revalidation found that overall predictive validity, per AUCs and odds ratios, remained stable relative to the 2021 study by Ferguson et al. (2021), with AUC scores ranging from 0.58 to 0.69, indicating fair to good predictive performance across all three outcome measures. However, Severson & Ferguson (2024) also identified some factors embedded within the PSA that had limited predictive power in the 2nd Judicial District, particularly age-related factors within the NCA and NVCA scales, highlighting the importance of ongoing monitoring and potential local adaptations (i.e., virtually no correlation between the factor and the outcome). Moreover, they similarly reported poor levels of predictive performance of the NVCA flag specifically, and speculate on some potential explanators for the attenuated predictive power of the NVCA flag in relation to the FTA and NCA scales (e.g., outcome rarity; binary nature of the flag versus a continuous scale). These findings underscore the dynamic nature of pretrial populations and court processes, highlighting the need for regular reevaluation and monitoring to ensure ongoing tool effectiveness and change in predictive quality over time in local populations. The 2nd Judicial District validation provides a foundation for understanding some of the potentially unique dynamics about the PSA's performance in New Mexico relative to other states.

Methods

To explore the association between scores on the PSA with pretrial outcomes, we acquired administrative data from the New Mexico Administrative Office of the Courts (AOC) encompassing all jurisdictions that implemented the PSA during the study period (i.e., 2019 – 2024). The PSA generates two primary risk scores: a Failure to Appear (FTA) scaled score ranging from 1-6 and a New Criminal Activity (NCA) scaled score ranging from 1-6, with higher scores indicating greater risk of the outcome (Lowenkamp et al., 2013). The PSA also generates a binary New Violent Criminal Activity (NVCA) flag, which is intended to spotlight defendants at a higher risk of engaging in violent new criminal activity if released pretrial. When reporting specifically on the outcomes – not the scale names, given that the names are internal to the tool – we refer to these outcomes hereafter as the court appearance rate (CAR) (as the inverse of the FTA rate), the public safety rate (as the inverse of the NCA rate), and the Nonviolence Rate (NVR) (as the inverse of the NVCA flag).

We restricted the sample to defendants who were released pretrial and had sufficient exposure time to realize potential outcomes. Specifically, we excluded cases if defendants remained in custody throughout

their case proceedings (e.g., if they were held on granted preventive detention motions) or if the time between release and case closure was less than three days, as insufficient exposure time could artificially deflate estimates of pretrial outcome rates and bias subsequent validation results. Given this, our final sample included defendants with complete outcome data and adequate exposure length, representing the population for whom pretrial risk assessment is most relevant to judicial decision-making.

We examined geographic variation in implementation and outcomes across three administrative levels: counties, judicial districts, and individual courts. We systematically analyzed patterns of missing demographic data across all administrative levels to assess potential data quality issues and their implications for testing for potential predictive biases in the PSA across demographic groups. We classified court appearance outcomes as either appeared or did not appear based on court records, while new criminal activity was defined as any new arrest or criminal charge filed during the pretrial period, consistent with standard pretrial outcome definitions used in validation research (Cohen & Reaves, 2007). We cross-referenced the AOC's reported NCA and NVCA outcome measures with manual review of jail booking records obtained from the relevant jail systems within the implementing jurisdictions.

Our descriptive analyses of the PSA's implementation consisted of reviewing the distribution of PSA risk scores across all implementing districts and statistically evaluating differences in baseline risk by geographic unit. However, for our core predictive validity analysis, we limited our sample to the 1st and 11th judicial districts, given higher confidence in the accuracy of jail and court data obtained from these districts (more on this in the *Analytic Sample for Predictive Validity* subsection), and explored how well the PSA FTA and NCA scaled scores, and NVCA flag, predicted pretrial outcomes.

Following Severson & Ferguson (2024), we assessed predictive validity of the PSA using complementary approaches, including receiver operating characteristic (ROC) analysis, logistic regression modeling, and calibration assessment across demographic subgroups. We calculated the area under the curve (AUC) statistics for both CAR and PSR risk scores at the court and district levels to evaluate discriminative accuracy (i.e., the PSA's ability to differentiate between a defendant who would or would not engage in new criminal activity at a given score level), with AUC values interpreted according to established benchmarks in the criminal justice discipline. We also fitted hierarchical logistic regression models to estimate predicted probabilities of pretrial outcomes across the full range of risk scores while controlling for demographic and legal factors, including age, sex, and supervision status. Our model specifications included mixed-effects models with random intercepts and slopes by judicial district, court, and year to account for clustering within geographic units and potential variation in risk-outcome relationships across jurisdictions (Gelman & Hill, 2006).

Moreover, we conducted differential validity testing to examine whether PSA predictive performance varied across demographic subgroups, with a particular focus on sex, to identify potential sources of bias or unfairness in risk prediction. We fitted separate logistic regression models for each demographic subgroup and compared predicted probability curves visually and statistically to assess whether risk scores demonstrated equivalent predictive relationships across groups. Our calibration assessment also explored whether predicted risk levels corresponded appropriately to observed outcome rates within demographic subgroups, providing evidence about whether the PSA performs fairly across different populations served by New Mexico's court system.

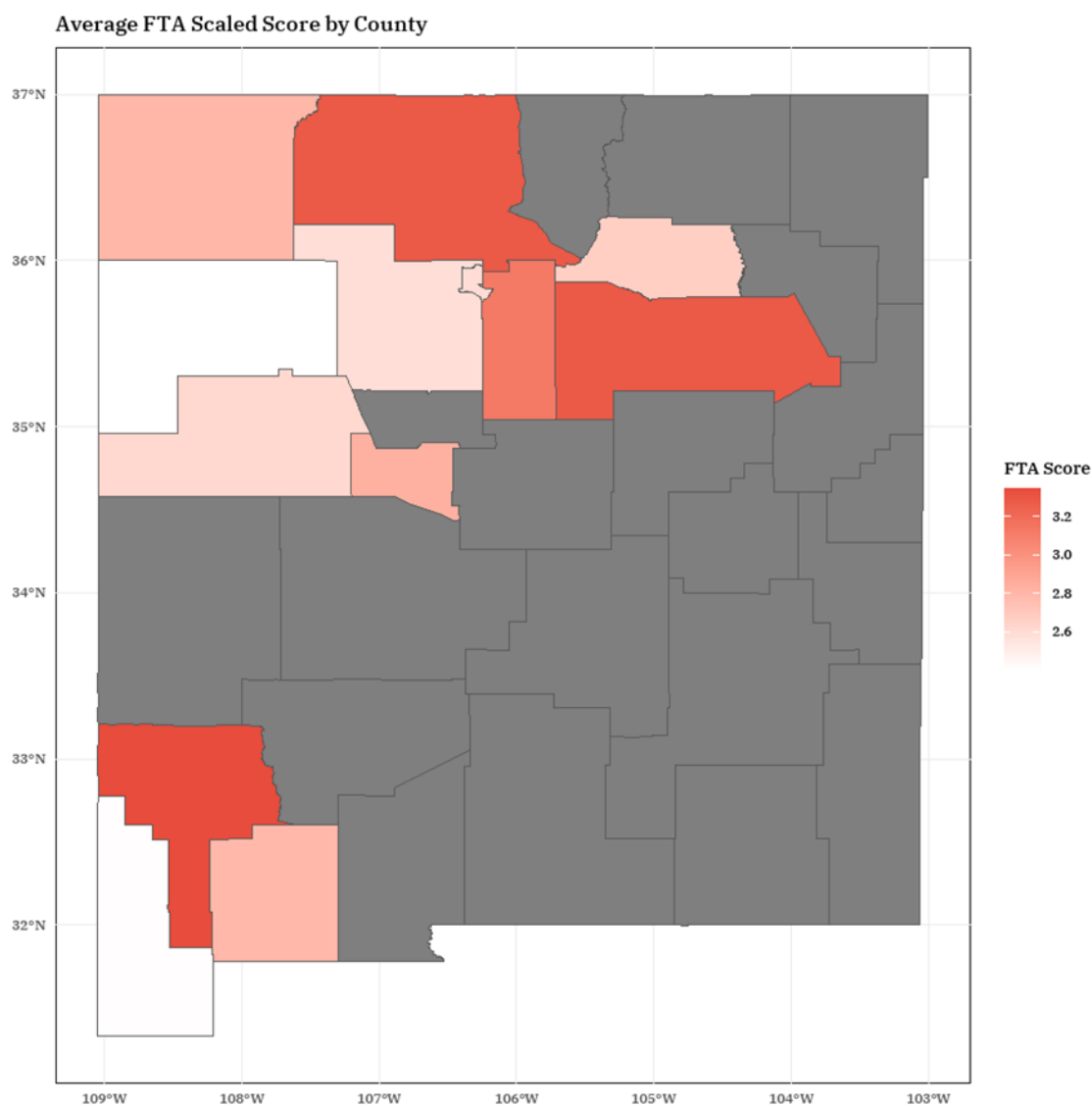
Results

Full Sample PSA Scores

We first present maps of average FTA, NCA, and NVCA scores – as well as supervision rates (i.e., the subset of sampled individuals with any PML designation) – by county in Figures 1 through 4 alongside descriptive statistics of PSA scores by county and judicial district, respectively, in Table 1 and Table 2. Per Figures 1-3 and Table 1, our analysis of PSA risk scores across New Mexico's implementing counties reveals variation in both sample sizes and risk score distributions, with San Juan County representing the largest jurisdiction (N = 13,651) and Los Alamos County the smallest (N = 74) at the time of data collection. FTA risk scores were geographically heterogeneous, ranging from a low of 2.41 ± 1.18 in Hidalgo County to a high of 3.35 ± 1.58 in Grant County, with median values consistently falling between 2 and 3

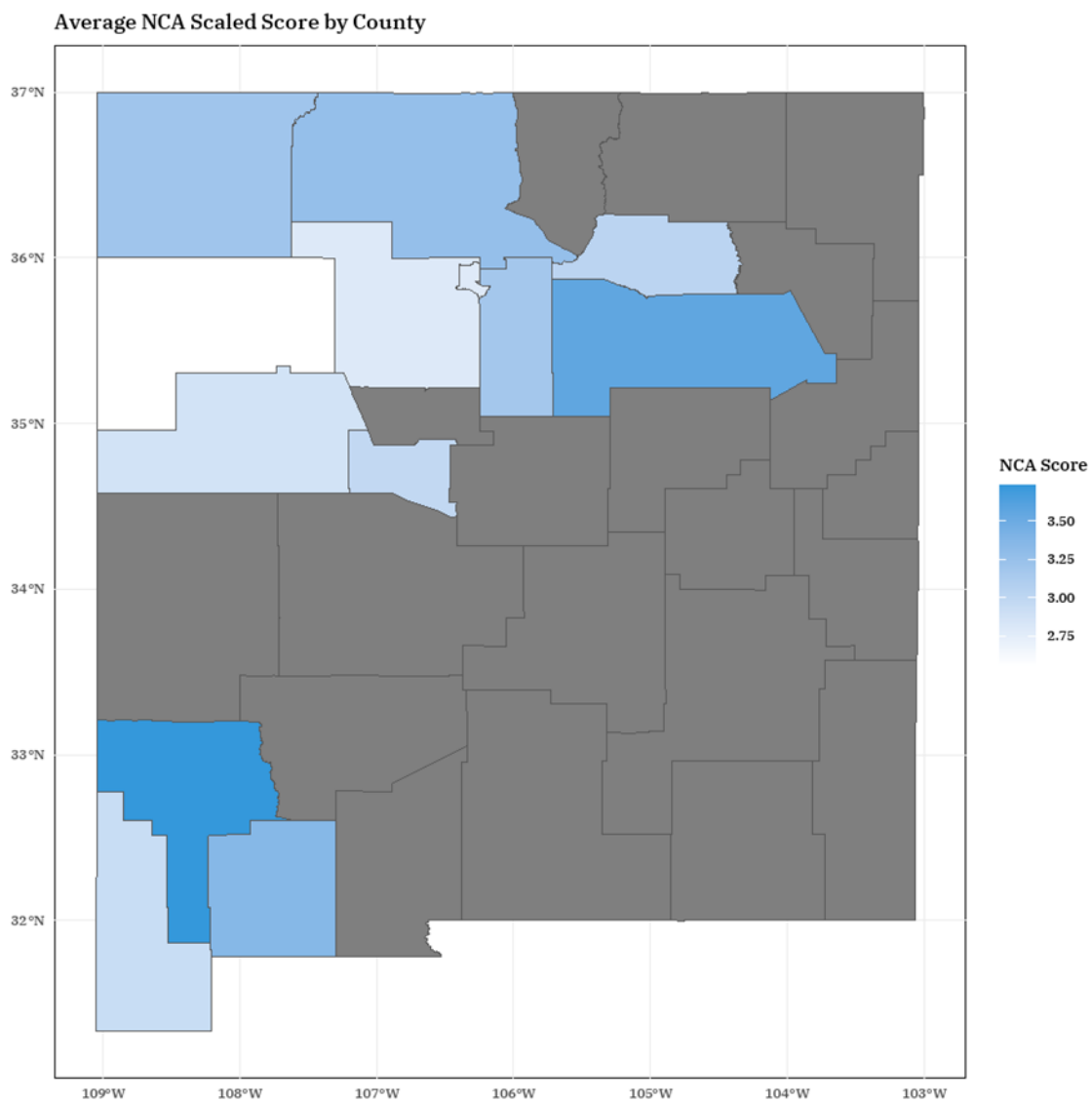
across most jurisdictions. There was greater inter-county variation in NCA scores, with Grant County displaying the highest mean score (3.74 ± 1.59) and McKinley County showing the lowest (2.56 ± 1.51), suggesting geographic differences in criminal recidivism risk that may reflect local demographic, socioeconomic, or criminal justice system characteristics. The NVCA flag also had notable county-level variation in the proportion of defendants identified as having violent crime risk factors, ranging from just 8% in Los Alamos County to 26% in Grant County, indicating that more than one in four defendants who were received a PSA assessment in Grant County were flagged for potential violent criminal activity compared to fewer than one in twelve in Los Alamos County. The pattern of higher risk scores across all three measures in Grant County, coupled with consistently lower scores in counties such as Los Alamos and Hidalgo, suggests that geographic location may serve as a proxy for underlying community characteristics that may, in turn, be associated with differences in pretrial risk assessment outcomes, with rural counties showing particularly diverse risk profiles that may warrant targeted implementation strategies and local contextual considerations in PSA score interpretation.

Figure 1. Average PSA FTA Risk Score by County



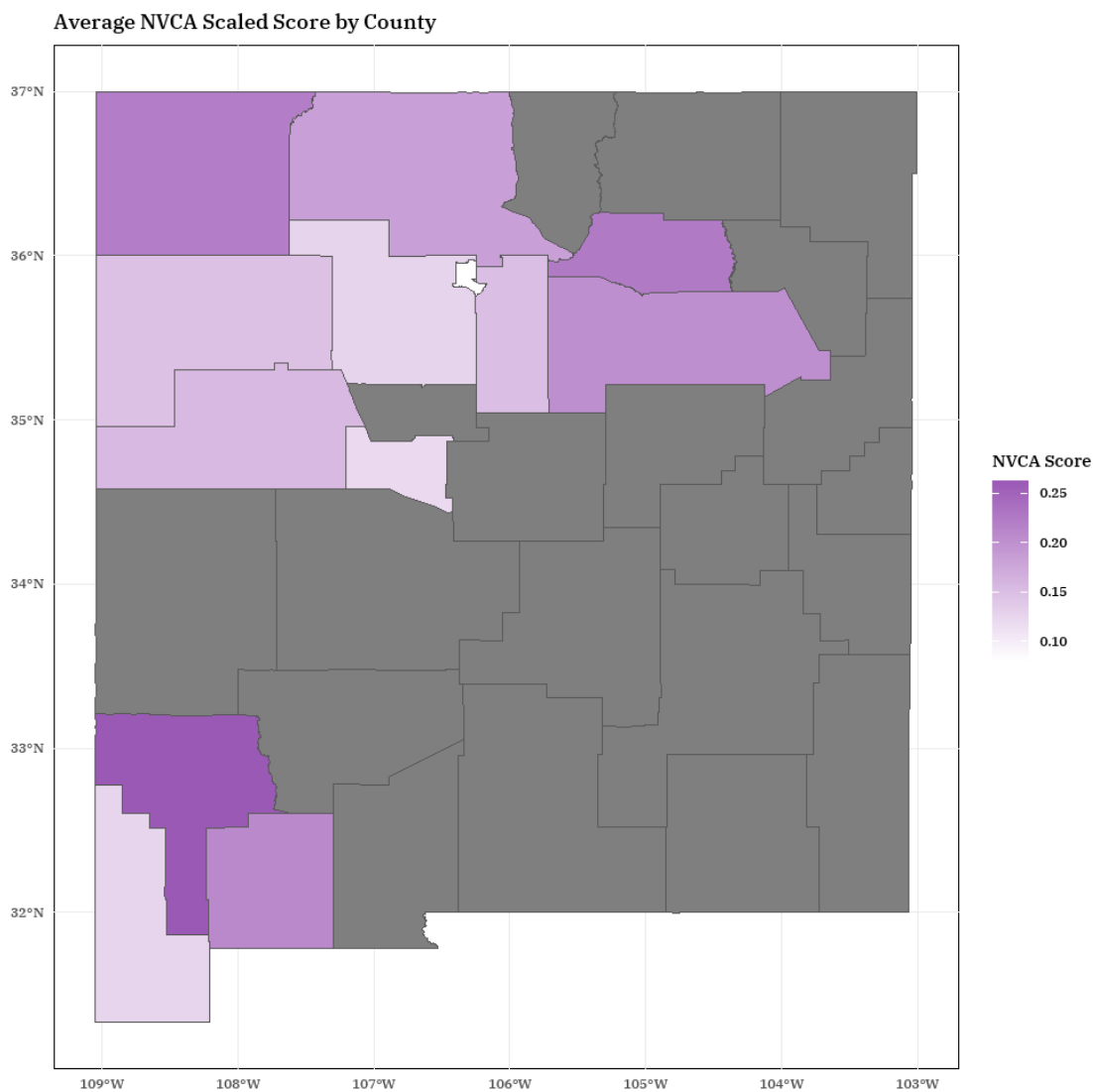
Data Source: County-level aggregation of failure to appear risk scores showing geographic variation in assessed risk levels (n = 53268).
Spatial distribution analysis provides insights into regional differences in pretrial population characteristics and risk profiles.
Analysis supports evidence-based resource allocation and targeted intervention strategies.
Geographic patterns inform understanding across New Mexico's diverse judicial landscape and demographic contexts.

Figure 2. Average PSA NCA Risk Score by County



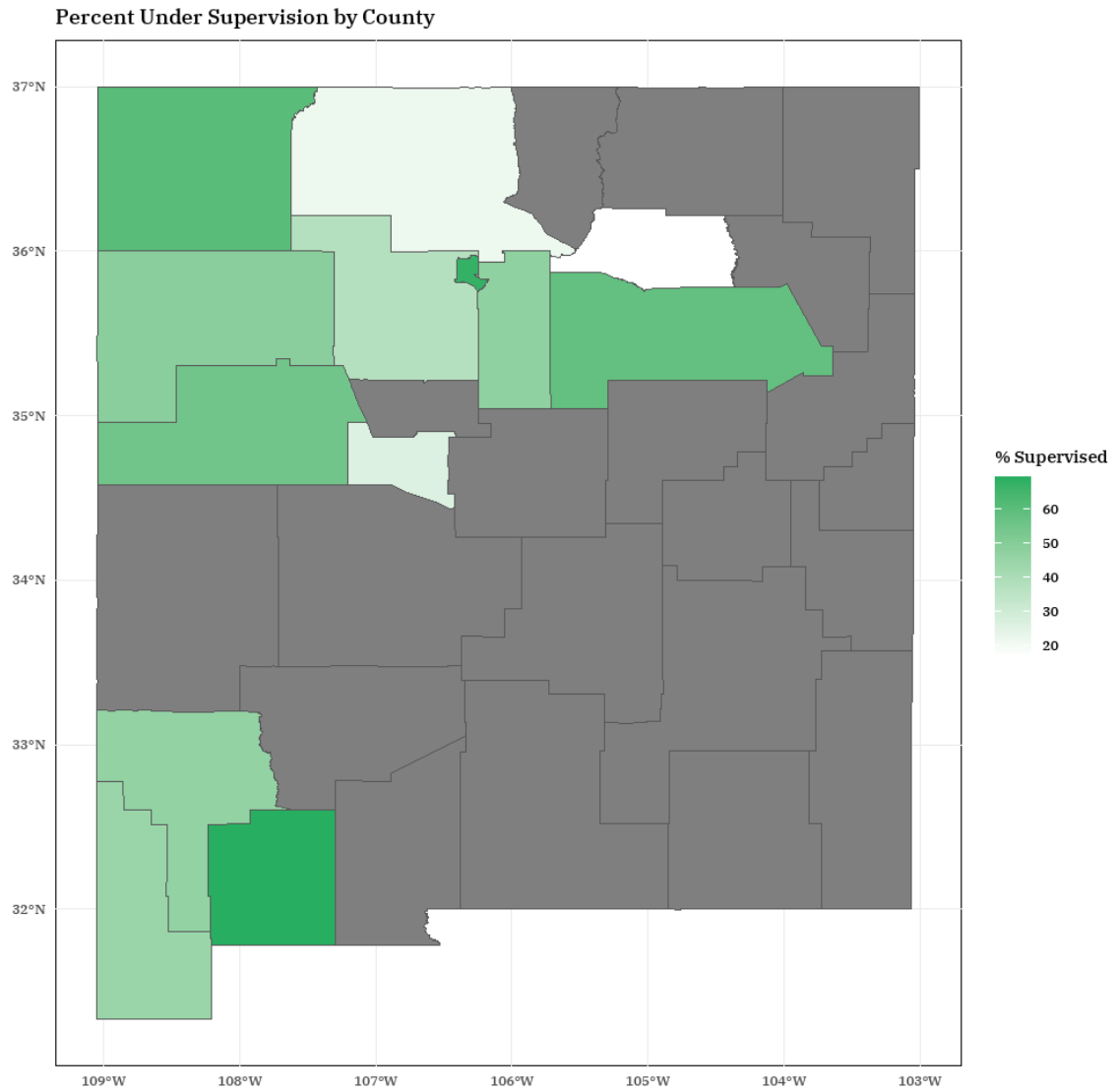
Data Source: Geographic analysis of new criminal activity risk scores across counties revealing spatial patterns in public safety risk assessment (n = 53268).
County-level variation provides critical information for understanding regional crime patterns.
Analysis informs targeted pretrial supervision strategies that account for local population characteristics.
Geographic patterns support evidence-based approaches across diverse criminal justice contexts.

Figure 3. Average PSA NVCA Risk Score by County



Data Source: Spatial distribution of new violent criminal activity risk scores demonstrating county-level variation in violent crime risk assessment (n = 53268).
Geographic analysis reveals important patterns in violent crime risk that inform targeted public safety interventions.
Analysis supports specialized supervision protocols and evidence-based pretrial decision-making.
Geographic patterns prioritize community safety while accounting for regional differences in violence risk factors.

Figure 4. *Percent Supervised by County*



Data Source: County-level supervision rates revealing geographic variation in pretrial monitoring practices (n = 53268).
Spatial analysis provides insights into differential implementation of supervision protocols across jurisdictions.
Analysis informs understanding of local practices and supports development of standardized guidelines.
Geographic patterns ensure equitable treatment while maintaining public safety objectives across diverse court systems.

Table 1. *PSA Risk Score Descriptive Statistics by County*

County	N	FTA Scaled Score	NCA Scaled Score	NVCA Scaled Score
San Juan	13,651	2.8 ± 1.39 (Med: 2)	3.2 ± 1.45 (Med: 3)	0.22 ± 0.42 (Med: 0)
Dona Ana	12,763	2.97 ± 1.61 (Med: 3)	3.29 ± 1.61 (Med: 3)	0.2 ± 0.4 (Med: 0)
Sandoval	7,676	2.58 ± 1.46 (Med: 2)	2.79 ± 1.46 (Med: 3)	0.13 ± 0.33 (Med: 0)
Santa Fe	4,958	3.13 ± 1.71 (Med: 3)	3.18 ± 1.61 (Med: 3)	0.15 ± 0.36 (Med: 0)
Valencia	3,413	2.84 ± 1.53 (Med: 3)	2.98 ± 1.48 (Med: 3)	0.12 ± 0.33 (Med: 0)
Grant	3,164	3.35 ± 1.58 (Med: 3)	3.74 ± 1.59 (Med: 4)	0.26 ± 0.44 (Med: 0)
McKinley	2,334	2.4 ± 1.31 (Med: 2)	2.56 ± 1.31 (Med: 2)	0.15 ± 0.35 (Med: 0)
Luna	1,543	2.79 ± 1.34 (Med: 3)	3.36 ± 1.47 (Med: 3)	0.21 ± 0.41 (Med: 0)
Rio Arriba	1,174	3.29 ± 1.66 (Med: 3)	3.27 ± 1.52 (Med: 3)	0.18 ± 0.39 (Med: 0)
San Miguel	1,156	3.29 ± 1.52 (Med: 3)	3.57 ± 1.53 (Med: 4)	0.2 ± 0.4 (Med: 0)
Cibola	848	2.62 ± 1.43 (Med: 2)	2.87 ± 1.44 (Med: 3)	0.16 ± 0.36 (Med: 0)
Hidalgo	408	2.41 ± 1.18 (Med: 2)	2.94 ± 1.36 (Med: 3)	0.13 ± 0.33 (Med: 0)
Mora	106	2.67 ± 1.21 (Med: 2)	3.03 ± 1.46 (Med: 3)	0.23 ± 0.42 (Med: 0)
Los Alamos	74	2.59 ± 1.52 (Med: 2)	2.78 ± 1.56 (Med: 2)	0.08 ± 0.27 (Med: 0)

The descriptive analysis of PSA risk scores across New Mexico's six PSA-implementing judicial districts¹ reveals variation in sample sizes and risk score distributions, with the Eleventh District representing the largest jurisdiction (N = 15,985) and the Fourth District the smallest (N = 1,262), likely due both to variable implementation timelines for PSA adoption and rollout as well as variation in population size (e.g., the Fourth District, consisting of Mora County and San Miguel County, had an implementation start date of June 2023 whereas the Eleventh District had an implementation start date of February 2020 for San Juan County and July 2022 for McKinley County). Failure to appear (FTA) risk scores demonstrate relatively modest inter-district variation, ranging from a low of 2.66 ± 1.48 in the Thirteenth District to a high of 3.24 ± 1.51 in the Fourth District, with median values consistently falling between 2 and 3 across all jurisdictions. New criminal activity (NCA) scores exhibit the greatest geographic heterogeneity among the three risk measures, with the Sixth District showing the highest mean score (3.56 ± 1.56) and the Thirteenth District displaying the lowest (2.85 ± 1.47), suggesting meaningful differences in assessed criminal recidivism risk across jurisdictions. The new violent criminal activity (NVCA) flag shows geographic variation in the proportion of defendants identified as having violent crime risk factors, ranging from 13% in the Thirteenth District to 24% in the Sixth District, indicating that nearly one in four

¹ While the PSA generates standardized risk scores across all implementing jurisdictions, the actual release conditions recommended for defendants are determined by locally-developed Decision-Making Frameworks (DMFs) or Release Conditions Matrices (RCMs) that translate PSA scores into specific supervision levels and release conditions. These frameworks vary between jurisdictions based on local policies, available resources, and stakeholder preferences. For example, Utah allows counties to customize their DMFs, with most using a standard framework but some implementing modified versions Decision-Making Frameworks (DMFs) by County. Similarly, San Francisco developed jurisdiction-specific overrides within their DMF for certain charges that increase supervision levels beyond what PSA scores alone would suggest. These local policy decisions mean that identical PSA scores may result in different release conditions across jurisdictions, reflecting community-specific approaches to balancing public safety, court appearance, and pretrial liberty while working within local resource constraints.

defendants in the Sixth District are flagged for potential violent criminal activity compared to roughly one in eight in the Thirteenth District. The consistency of median values near the lower end of score ranges for FTA and NCA measures across districts suggests that while mean scores may differ geographically, the majority of defendants across all jurisdictions receive relatively low risk scores, with district-level differences primarily driven by variations in the distribution of higher-risk cases and the prevalence of violent crime risk factors rather than systematic shifts in the entire risk profile.

Table 2. *PSA Risk Score Descriptive Statistics by Judicial District (Mean and Median)*

Judicial District	N	FTA Scaled Score	NCA Scaled Score	NVCA Scaled Score
Eleventh	15,985	2.74 ± 1.39 (Med: 2)	3.11 ± 1.45 (Med: 3)	0.21 ± 0.41 (Med: 0)
First	6,206	3.15 ± 1.7 (Med: 3)	3.19 ± 1.59 (Med: 3)	0.16 ± 0.36 (Med: 0)
Fourth	1,262	3.24 ± 1.51 (Med: 3)	3.52 ± 1.53 (Med: 4)	0.2 ± 0.4 (Med: 0)
Sixth	5,115	3.1 ± 1.52 (Med: 3)	3.56 ± 1.56 (Med: 4)	0.24 ± 0.42 (Med: 0)
Third	12,763	2.97 ± 1.61 (Med: 3)	3.29 ± 1.61 (Med: 3)	0.2 ± 0.4 (Med: 0)
Thirteenth	11,937	2.66 ± 1.48 (Med: 2)	2.85 ± 1.47 (Med: 3)	0.13 ± 0.33 (Med: 0)

Per Table 3, the formal statistical tests of geographic differences in PSA risk scores suggest statistically significant variation across both county and district jurisdictions in New Mexico, with all ANOVA tests yielding highly significant results ($p < 0.001$) for FTA, NCA, and NVCA measures. While the statistical significance indicates that geographic differences exist beyond what would be expected by chance alone, the practical magnitude of these differences is relatively modest, as evidenced by small effect sizes across all measures (η^2 ranging from 0.009 to 0.029). The NCA risk scores demonstrate the largest geographic variation at both county (eta-squared, $\eta^2 = 0.029$) and district (eta-squared, $\eta^2 = 0.019$) levels, suggesting that local contextual factors may have a slightly greater influence on new criminal activity risk assessment compared to failure to appear or violent crime risk. District-level analyses show higher F-statistics than county-level comparisons, reflecting the reduced number of comparison groups (6 districts versus 14 counties), though the effect sizes at the district level are consistently smaller, indicating that while differences in baseline risks between districts are statistically robust, they account for a smaller proportion of total variance in risk scores.

To put these effect sizes into perspective, we can think of them in terms of how much geographic location actually matters for predicting PSA risk scores, with the eta-squared (η^2) values ranging from 0.009 to 0.029, indicating that geographic location explains only about 1% to 3% of the total variation in risk scores across individuals. Using Cohen's conventional benchmarks for interpreting effect sizes, these values fall into the "small" effect size category, with most clustering around the lower end of what researchers typically consider meaningful. Even the largest effect size ($\eta^2 = 0.029$ for county-level NCA scores) suggests that knowing which county someone is from would improve our ability to predict their risk score by less than 3% compared to not knowing their location at all. The remaining 97-99% of the variation in risk scores appears to stem from individual-level characteristics rather than where someone lives within New Mexico, suggesting that while we can detect geographic differences, these differences may not translate into meaningful disparities in how the PSA tool functions across different areas of the state, at least conditional on baseline risk. This pattern suggests that individual-level factors are likely stronger predictors of pretrial outcomes than contextual or geographic variables and that geographic location contributes a statistically detectable, but practically small, amount of variation in PSA risk scores.

Table 3. *Statistical Tests for Geographic Differences in PSA Risk Scores*

Geographic Level	PSA Score Type	Number of Groups	F-statistic	p-value	η^2 Effect Size
County	FTA Scaled Score	14	98.656	< 0.001	0.024
County	NCA Scaled Score	14	123.000	< 0.001	0.029
County	NVCA Scaled Score	14	47.733	< 0.001	0.012
District	FTA Scaled Score	6	160.485	< 0.001	0.015
District	NCA Scaled Score	6	205.760	< 0.001	0.019
District	NVCA Scaled Score	6	94.957	< 0.001	0.009

Analytic Sample for Predictive Validity

From the full dataset of 53,268 cases we received from the AOC, we only retained 20.4% ($n = 10,872$) of cases for outcome-specific analyses of predictive validity across all three pretrial outcomes. This 79.6% exclusion rate primarily reflects limiting cases to those with complete outcome data and adequate exposure periods, as well as excluding data from multiple judicial districts due to unreliable reported jail data, which decreases confidence in the observed associations between PSA scale scores and outcomes. For more on the scope of agreement between AOC's records of FTA, NCA, and NVCA outcomes and those manually evaluated by CARA staff in the analytic sample from the 1st and 13th judicial districts, see Appendix A.

More specifically, regarding the exclusion of other districts, we found that jail data quality issues were pervasive across jurisdictions, with release log completeness varying substantially by county. Our analysis revealed that missing release logs affected at least 40.5% of cases overall, though we suspect this likely underestimates the true scope of missingness, given that some jurisdictions maintain multiple daily logs across different shifts. We observed that individual counties experienced missing release log rates of 50-60%, which appear to reflect both systematic reporting gaps and inconsistent data-collection practices across facilities. Even in Santa Fe County, which showed the lowest percentage of missing release logs (i.e., less than 1% missing), we identified hundreds of cases with errors such as consistent date discrepancies or incomplete entries, suggesting that data quality issues persist even when overall reporting rates at the level of the release log appear sufficient. Moreover, the form of documentation also complicated data extraction efforts, as many release logs were provided as handwritten or poorly scanned PDF documents that proved incompatible with automated data processing methods, inclusive of optical character recognition (OCR) scanning in *R*. These systematic data quality challenges contributed to our decision to exclude multiple judicial districts from the analysis given a lack of confidence in the reported data, though we recognize that this limits the geographic coverage of the predictive validity testing.

Additionally, even within the retained analytic sample, there was nontrivial missingness in race and ethnicity data, affecting 64.6% of the sample (see Table 4 and Figure 5). This limits our capacity to conduct sufficiently statistically powered analyses examining potential disparities in PSA tool performance across racial and ethnic subgroups, given small cell sizes (i.e., a small number of individuals of specific race-ethnicity categories represented across all six scale score categories).

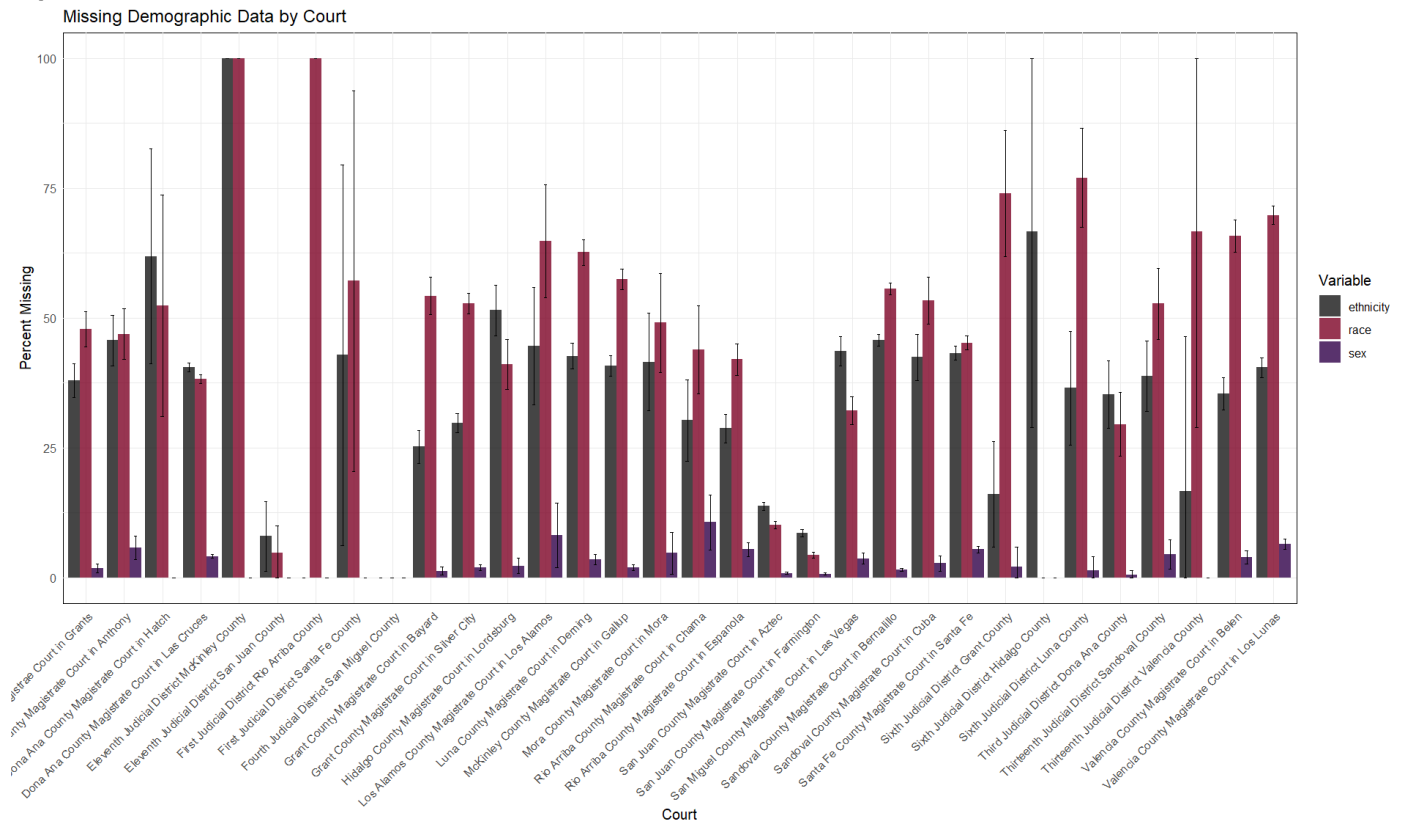
Having noted these exclusion criteria and limitations, we present our primary descriptive statistics. First, we examined the sociodemographic profile of the sample, which included 10,871 defendants with complete outcome data. The sample had a mean age of 35.8 years ($SD = 11.4$), with the largest age group being 25-34 years at 35.0% ($n = 3,800$), followed by 35-44 years at 28.3% ($n = 3,074$), 45+ years at 20.4%

(n = 2,214), and 18-24 years at 16.4% (n = 1,778). Men comprised 73.7% (n = 7,762) of the sample. Race and ethnicity data revealed substantial missing information, with 64.6% (n = 7,019) classified as other/unknown, while Hispanic defendants represented 23.3% (n = 2,528), White defendants 8.8% (n = 958), American Indian defendants 2.5% (n = 275), Black defendants 0.7% (n = 78), Asian defendants 0.1% (n = 8), and Pacific Islander defendants 0.0% (n = 5).

Table 4. *Sociodemographic Characteristics of the Analysis Sample*

Characteristic	Analytic Sample (N = 10,872)
N	10,872
Age, mean (SD)	35.77 (11.38)
Age Category (%)	
18-24	1,778 (16.4%)
25-34	3,800 (35.0%)
35-44	3,074 (28.3%)
45+	2,215 (20.4%)
Unknown	5 (0.0%)
Sex, Male (%)	7,763 (73.8%)
Race-Ethnicity (%)	
American Indian	275 (2.5%)
Asian	8 (0.1%)
Black	78 (0.7%)
Hispanic	2,528 (23.3%)
Missing	7,020 (64.6%)
Pacific Islander	5 (0.0%)
White	958 (8.8%)

Figure 5. Missing Demographic Data by Court



Data Source: Court-specific examination of demographic data quality patterns revealing heterogeneity in data collection practices and completeness across individual court jurisdictions. Error bars demonstrate statistical precision of missingness estimate

Per Table 5, within our sample, average PSA risk scores were 2.68 (SD = 1.52) for FTA, 2.87 (SD = 1.49) for NCA, and 13% of the sample had NVCA flags. Pretrial supervision restrictions (i.e., PML 1 – 4) were ordered for 39.4% (n = 4,280) of defendants. The mean exposure length of the subset of defendants who were exposed was 114.5 days (SD = 109). Outcome rates were 76.0% (n = 8,263) for CAR, 80.0% (n = 8,698) for PSR, and 94.0% (n = 10,220) for NVR. Geographic representation was heavily concentrated in the 13th Judicial District, which represented 75.9% (n = 8,253) of the sample.

Table 5. Case and PSA Characteristics of the Analytic Sample

Characteristic	Analytic Sample (N = 10,871)
N	10,871
FTA Scaled Score, mean (SD)	2.68 (1.52)
NCA Scaled Score, mean (SD)	2.87 (1.49)
NVCA Scaled Score, mean (SD)	0.13 (0.33)
Pretrial Supervision, Yes (%)	4280 (39.4)
Released Pretrial, Yes (%)	9067 (83.4)
Exposure Days, mean (SD)	114.45 (109.05)
Court Appearance Rate (CAR) (mean (SD))	0.76 (0.43)
Public Safety Rate (PSR) (mean (SD))	0.80 (0.40)
Non-Violence Rate (NVR) (mean (SD))	0.94 (0.23)
Judicial District, Thirteenth (%)	8,253 (75.9)

District and Court-Level Summaries

The district-level analysis reveals statistically significant differences in pretrial outcome rates between the First and Thirteenth judicial districts across all three pretrial outcome measures. The Thirteenth District had a Court Appearance Rate of 78.9% compared to the First District's 65.0%, representing a 13.9 percentage point difference. Public Safety Rates also differed between districts, with the Thirteenth District at 82.0% versus the First District's 75.9%, indicating a 6.1 percentage point difference in new criminal activity during the pretrial period. Both districts maintained high Non-Violence Rates, with the Thirteenth District at 94.4% and the First District at 93.8%, a 0.6 percentage point difference in pretrial violent behavior. The difference in sample sizes, with the Thirteenth District encompassing 8,253 cases compared to the First District's approximately 2,619 cases, reflects the jurisdictional scope and caseload differences between these two judicial districts. It is important to note that this analysis of outcome rates does not statistically account for other factors that may influence these rates, such as demographic characteristics, offense types, case complexity, resource availability, or implementation practices, which could contribute to the observed differences between jurisdictions and should be considered when interpreting these comparative findings.

Table 6. *District-Level CAR, PSR, and NVR Rates*

Outcome Measure	First District Sample Size	First District Rate (%)	Thirteenth District Sample Size	Thirteenth District Rate (%)
Court Appearance Rate (CAR)	2,618	65.0	8,253	78.9
Community Safety Rate (PSR)	2,619	75.9	8,253	82.0
Non-Violence Rate (NVR)	2,619	93.8	8,253	94.4

Similarly, from Table 7, our court-level analysis reveals variation in pretrial outcome rates across five major courts within the First and Thirteenth judicial districts. CARs ranged from 65.1% at Santa Fe County Magistrate Court in Santa Fe to 81.7% at Valencia County Magistrate Court in Los Lunas, representing a 16.6 percentage point difference across jurisdictions. Similarly, the PSR ranged from 75.9% at Santa Fe County Magistrate Court in Santa Fe to 84.8% at Valencia County Magistrate Court in Belen, a range of 8.9 percentage points. NVRs remained consistently high across all courts, ranging from 93.8% at Santa Fe County Magistrate Court in Santa Fe to 96.2% at Valencia County Magistrate Court in Belen, with only 2.4 percentage points separating the highest and lowest performing courts. Sample sizes vary considerably across jurisdictions, ranging from 373 cases at the Sandoval County Magistrate Court in Cuba to 5,565 cases at the Sandoval County Magistrate Court in Bernalillo, reflecting differences in caseload volume, jurisdictional scope, and variable PSA implementation start dates.

Table 7. County-Level CAR, PSR, and NVRs

Court	District	Outcome	Sample Size	Rate (%)
Sandoval County Magistrate Court in Bernalillo	Thirteenth	CAR	5,565	79.4
Sandoval County Magistrate Court in Bernalillo	Thirteenth	PSR	5,565	81.4
Sandoval County Magistrate Court in Bernalillo	Thirteenth	NVR	5,565	94.0
Santa Fe County Magistrate Court in Santa Fe	First	CAR	2,617	65.1
Santa Fe County Magistrate Court in Santa Fe	First	PSR	2,618	75.9
Santa Fe County Magistrate Court in Santa Fe	First	NVR	2,618	93.8
Valencia County Magistrate Court in Los Lunas	Thirteenth	CAR	1,654	81.7
Valencia County Magistrate Court in Los Lunas	Thirteenth	PSR	1,654	83.3
Valencia County Magistrate Court in Los Lunas	Thirteenth	NVR	1,654	95.7
Valencia County Magistrate Court in Belen	Thirteenth	CAR	584	76.0
Valencia County Magistrate Court in Belen	Thirteenth	PSR	584	84.8
Valencia County Magistrate Court in Belen	Thirteenth	NVR	584	96.2
Sandoval County Magistrate Court in Cuba	Thirteenth	CAR	373	72.9
Sandoval County Magistrate Court in Cuba	Thirteenth	PSR	373	82.3
Sandoval County Magistrate Court in Cuba	Thirteenth	NVR	373	93.3

Predictive Validity – FTA, NCA, and NVCA Scores and Observed Rates

The first step to evaluating the predictive validity of the PSA involves exploring whether empirical outcome rates align with theoretical expectations based on risk score classifications. For the PSA, we would expect to observe monotonic (i.e., strictly declining) relationships between risk scores and corresponding outcome rates, such that higher risk scores on a subscale correspond to higher rates of the negative outcome being predicted (or conversely, lower rates of the positive outcome). A monotonic relationship indicates that the rates consistently move in one direction (either increasing or decreasing) across risk levels without reversals or irregular patterns. This type of analysis provides initial evidence on whether the PSA successfully differentiates between defendants based on their likelihood of engaging in the predicted behavior.

We present rates of court appearance, public safety, and non-violence by FTA, NCA, and NVCA flags in Tables 8-10. The CAR analysis reveals a consistent inverse relationship between FTA risk scores and court appearance rates, supporting the predictive validity of the FTA scale. CARs declined progressively from 90.0% for defendants with FTA Score 1 (95% CI: 88.9%-91.1%) to 48.5% for those with FTA Score 6 (95% CI: 44.8%-52.1%), representing a statistically significant 41.5 percentage point difference across the risk spectrum (p -value < 0.001). The monotonic decrease in appearance rates across all six risk levels, with each successive score category having a lower appearance rate than the previous level, indicates that the FTA score, as given by the PSA, effectively stratifies defendants according to their likelihood of appearing for court proceedings.

Table 8. *Observed Court Appearance Rates by PSA FTA Risk Score*

FTA Score	Sample Size	CAR (%)	95% CI
1	2,875	90.0	(88.9%, 91.1%)
2	3,069	83.2	(81.9%, 84.5%)
3	2,013	72.8	(70.8%, 74.7%)
4	1,200	63.3	(60.6%, 66.1%)
5	1,004	50.0	(46.9%, 53.1%)
6	710	48.5	(44.8%, 52.1%)

Similarly, the PSR analysis supports the initial predictive validity of the PSA's NCA scale, with rates declining consistently across increasing risk score levels. PSRs decreased from 92.0% for defendants with NCA Score 1 (95% CI: 90.9%-93.1%) to 59.6% for those with NCA Score 6 (95% CI: 56.1%-63%), representing a statistically significant 32.4 percentage point difference between the lowest and highest risk categories (p-value < 0.001). The pattern across all six risk levels, where each higher score corresponds to a lower community safety rate, indicates that the NCA instrument effectively differentiates defendants based on their likelihood of remaining free from NCA during the pretrial period.

Table 9. *Observed Community Safety Rate by PSA NCA Risk Score*

NCA Score	Sample Size	PSR Rate (%)	95% CI
1	2,310	92.0	(90.9%, 93.1%)
2	2,825	87.8	(86.6%, 89%)
3	2,158	79.7	(78%, 81.4%)
4	1,940	71.9	(69.9%, 73.9%)
5	867	66.0	(62.8%, 69.1%)
6	772	59.6	(56.1%, 63%)

Additionally, our analysis of the NVR by NVCA flag status supports the predictive validity of the binary violence flag, with defendants flagged for elevated violent crime risk exhibiting lower non-violence rates than those without the NVCA flag. NVRs were 95.3% for defendants without an NVCA flag (95% CI: 94.9%-95.7%) compared to 87.3% for those with a flag present (95% CI: 85.6%-89.1%), representing an 8.0 percentage point difference between the two groups. This separation between flagged and non-flagged defendants suggests that the NVCA flag identifies individuals with elevated risk for violent criminal behavior during the pretrial period.

Table 10. *Observed Nonviolence Rate by PSA NVCA Risk Flag*

NVCA Flag Status	Sample Size	NVR Rate (%)	95% CI
No Flag	9,484	95.3	(94.9%, 95.7%)
Flag Present	1,388	87.3	(85.6%, 89.1%)

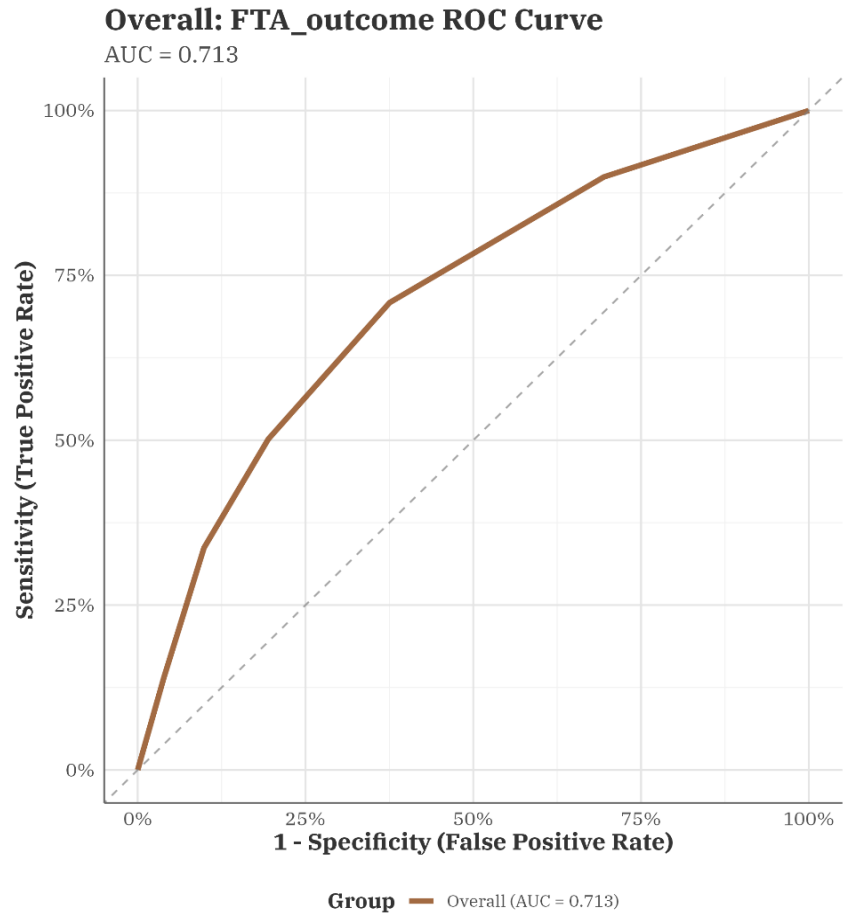
Predictive Validity – Full Sample AUC Results

Area Under the Curve (AUC) analysis represents a more rigorous approach to evaluating predictive validity compared to examining observed outcome rates by risk score levels. While raw outcome rates by score category can suggest risk gradients, this approach presents several methodological limitations that may obscure the true predictive performance of the PSA. Unadjusted outcome rates fail to account for the distributional properties of scores within each category, potentially masking important variations in predictive accuracy across different threshold values. Simple rate comparisons also cannot adequately capture the trade-offs between sensitivity and specificity that occur as decision thresholds change across the scoring continuum. The AUC methodology addresses these limitations by providing a single summary statistic that quantifies discriminative performance across all possible decision thresholds, enabling direct statistical comparisons between different geographies, population subgroups, and implementation contexts. This approach accounts for the full range of classification performance characteristics, offering a more nuanced understanding of how well risk scores separate individuals who experience pretrial failure from those who do not across the entire distribution of risk predictions.

The interpretation of AUC values in criminal justice risk assessment applications follows established benchmarks outlined by Desmarais and Singh (2013), who proposed that AUC values between 0.56 and 0.63 represent small effect sizes with poor levels of predictive performance, values between 0.64 and 0.70 indicate medium effect sizes with fair predictive accuracy, and values above 0.71 demonstrate large effect sizes with good to excellent discriminative ability. These standards reflect the recognition that perfect prediction (AUC = 1.0) is unrealistic in criminal justice applications, given the complex and often noisy nature of human behavior, while values at or near chance levels (AUC = 0.50) indicate the absence of meaningful predictive validity (i.e., that the PSA performs as well as a coin flip).

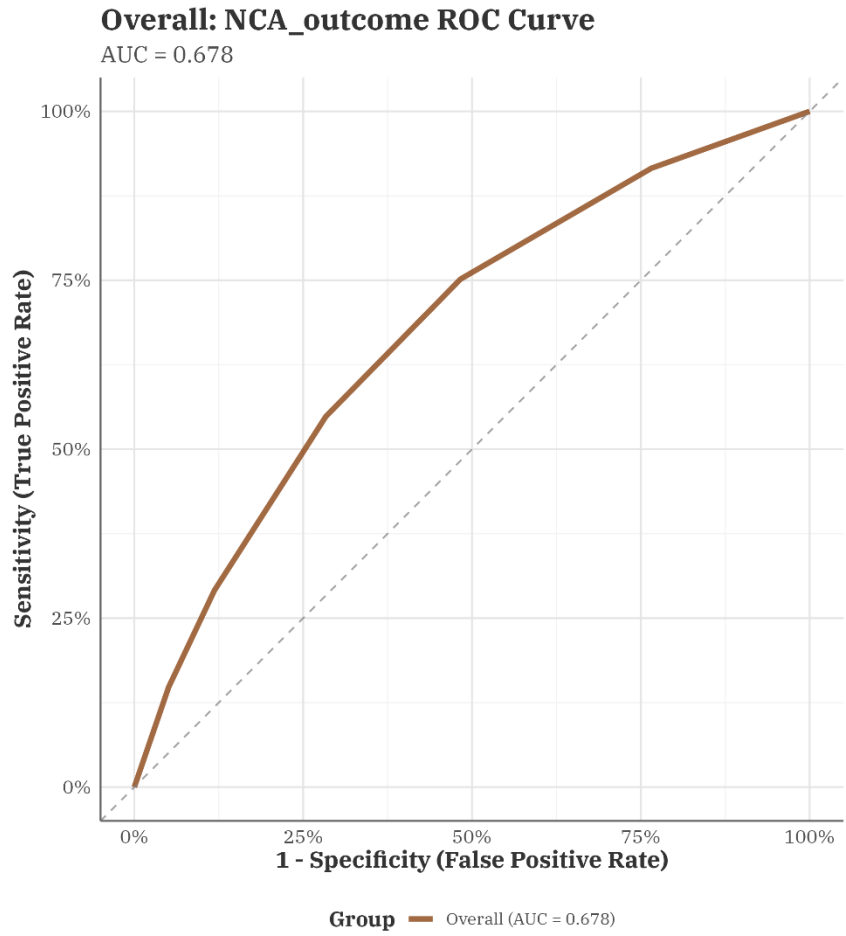
Our AUC analysis reveals differential predictive performance across the three PSA risk components, with discriminative ability varying substantially by outcome type. The FTA scale demonstrated the strongest predictive validity with an AUC of 0.713, achieving the large effect size threshold established by Desmarais and Singh (2013), indicative of good discriminative ability. The NCA scale demonstrated acceptable performance with an AUC of 0.678, falling within the medium effect size range, which represents fair predictive accuracy. However, the NVCA flag yielded an AUC of 0.584, which falls within the small effect size range and approaches the threshold for limited practical utility (i.e., poor predictive performance), suggesting that while the violent crime flag provides some discriminative capacity above chance levels, its predictive strength is considerably weaker than the PSA's FTA and NCA subscales. This pattern of results suggests that the PSA demonstrates reasonable performance in predicting court appearance and general criminal activity outcomes, yet has less accurate predictions of NVCA, which we suggest may arise from the combination of the rarity of the outcome being observed (i.e., violence is more rare than nonviolence), the higher noise-volume of the outcome making it inherently more challenging to predict at the individual-level, and the binary, versus continuous, nature of the flag which increases risk of classification error just as a numerical property of the flag versus a scale. We note that these broad results are roughly consistent with the AUC ranges we reported on in the initial validation and revalidation studies of the PSA in the Second Judicial District (Ferguson et al., 2021; Severson & Ferguson, 2024).

Figure 5. ROC Curve for PSA FTA Scores



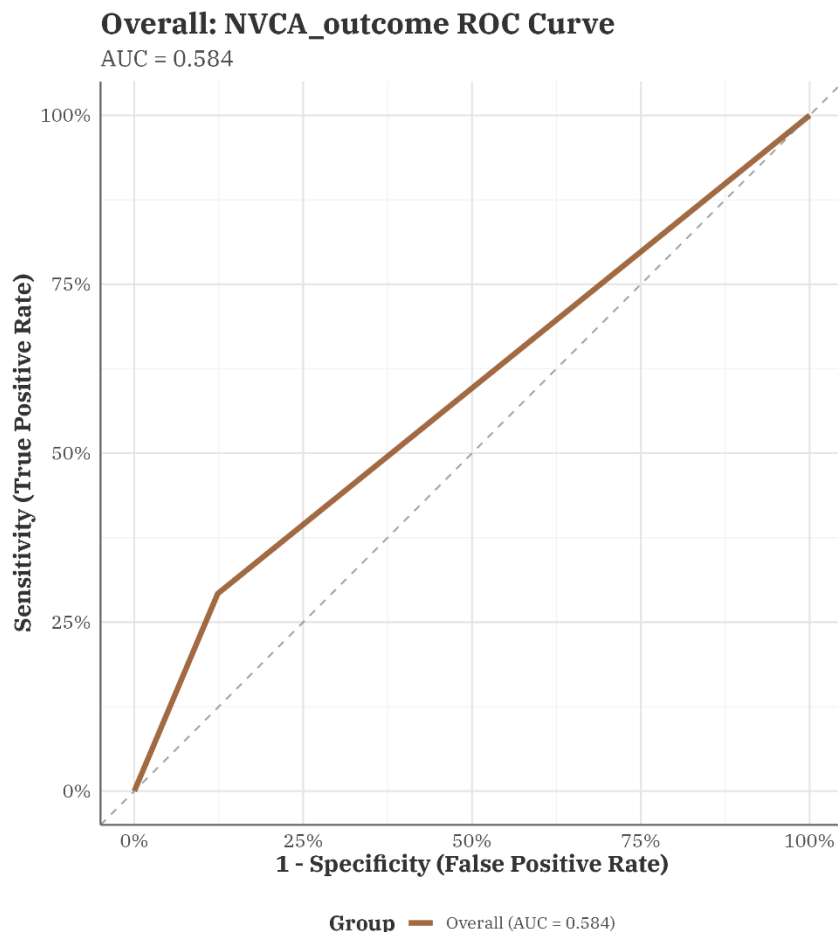
ROC curve demonstrates discriminative performance of FTA risk scores in predicting failure to appear outcomes across First and Thirteenth judicial districts. Area Under the Curve (AUC) values above 0.7 indicate good discriminative ability for evidence-based pretrial decision-making.

Figure 6. ROC Curve for PSA NCA Scores



ROC curve evaluates discriminative performance of NCA risk scores in predicting new criminal activity during pretrial periods. Performance assessment supports evidence-based pretrial supervision and intervention decisions.

Figure 7. ROC Curve for PSA NVCA Scores

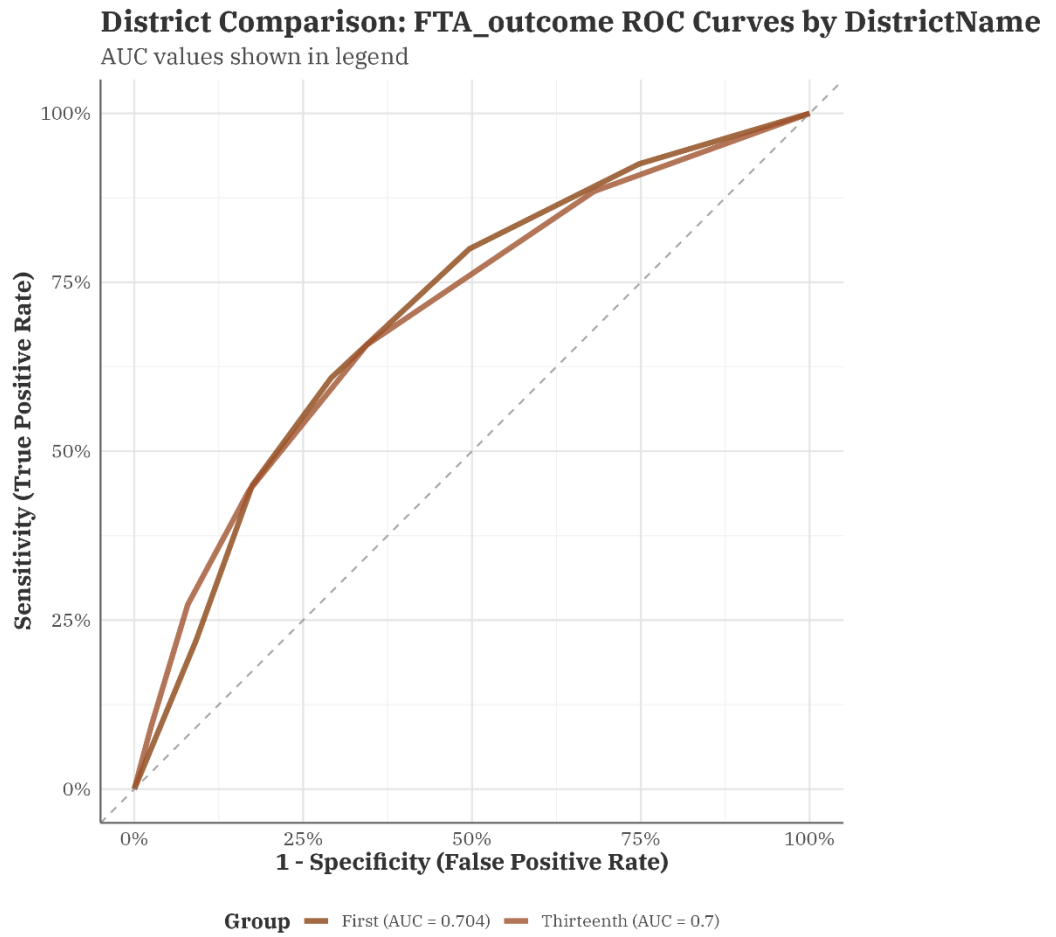


ROC curve demonstrates predictive performance of NVCA risk scores for new violent criminal activity outcomes. This high-stakes outcome assessment is critical for public safety-focused pretrial decisions.

Predictive Validity - District-Level AUC Results

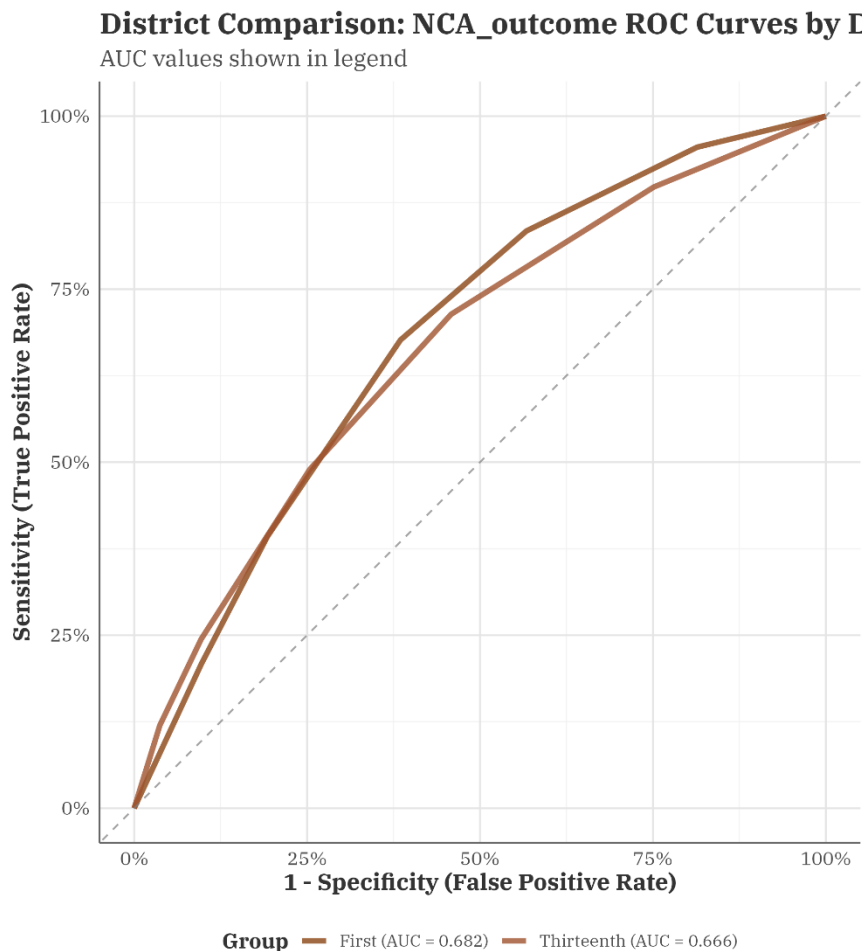
The district-level AUC analysis reveals consistent cross-jurisdictional performance with subtle but meaningful variations between the First and Thirteenth judicial districts across all three PSA risk scales or flags. For the CAR, the FTA subscale has good predictive validity within the large effect size range, with the First District achieving an AUC of 0.704 and the Thirteenth District achieving 0.700, indicating virtually equivalent discriminative ability for predicting court appearance outcomes across districts. For the PSR, the NCA subscale has acceptable levels of predictive validity in both jurisdictions, with the First District (AUC = 0.682) slightly outperforming the Thirteenth District (AUC = 0.666), though these differences in AUCs are not statistically significant. We observe the most pronounced district-level difference in NVR prediction from the NVCA flag, where the flag for the First District has an AUC of 0.611 compared to 0.574 in the Thirteenth District. This suggests that while the PSA generally has reasonable predictive accuracy at predicting court appearance and general criminal activity prediction across judicial districts, the identification of violent crime risk may be more sensitive to local population characteristics, implementation practices, or contextual factors that vary between districts, highlighting the importance of jurisdiction-specific validation for high-stakes public safety decisions involving potential violent crime risk assessment.

Figure 8. District-Specific ROC Curve for PSA FTA Scores



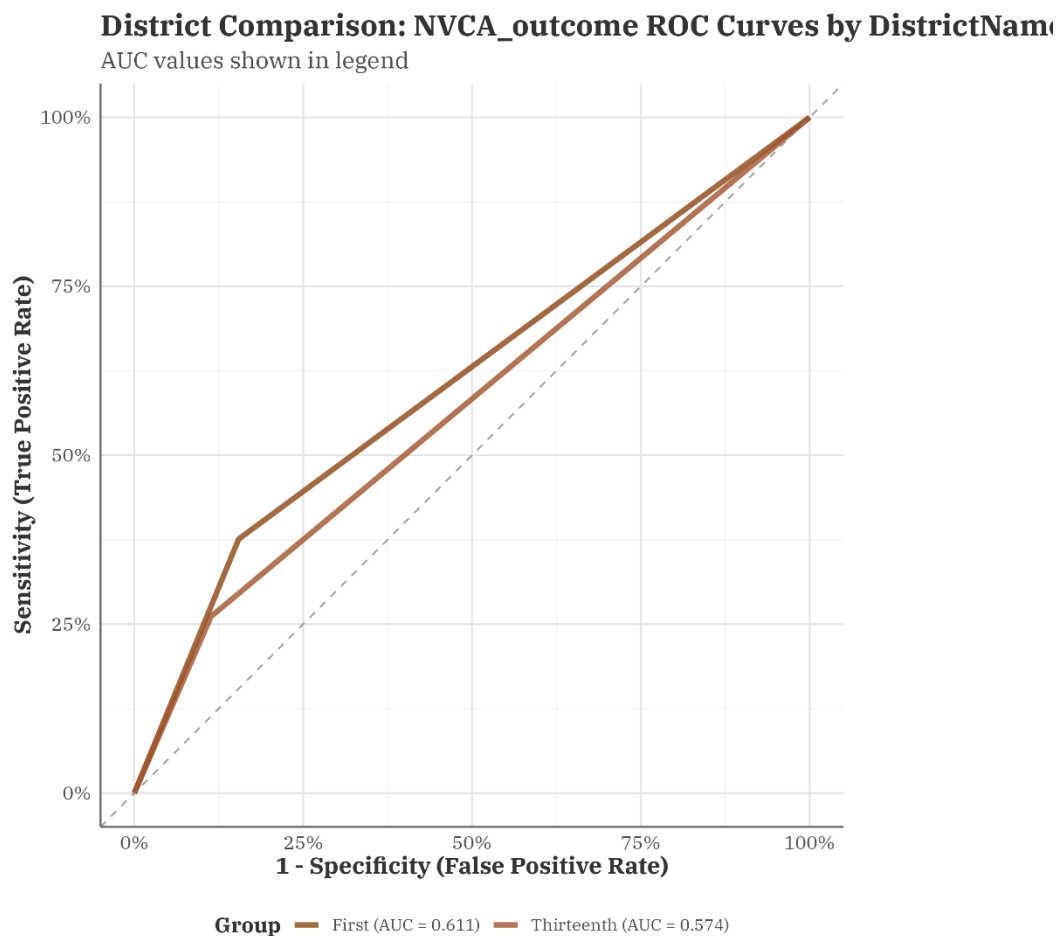
District-specific ROC curves compare FTA risk score performance between First and Thirteenth judicial districts, revealing jurisdictional differences in risk prediction effectiveness.

Figure 9. District-Specific ROC Curve for PSA NCA Scores



District-level ROC curve comparison assesses NCA risk score consistency across First and Thirteenth judicial districts.

Figure 10. District-Specific ROC Curve for PSA NVCA Flag



District-specific NVCA risk score performance comparison provides critical insights for public safety-focused pretrial decisions.

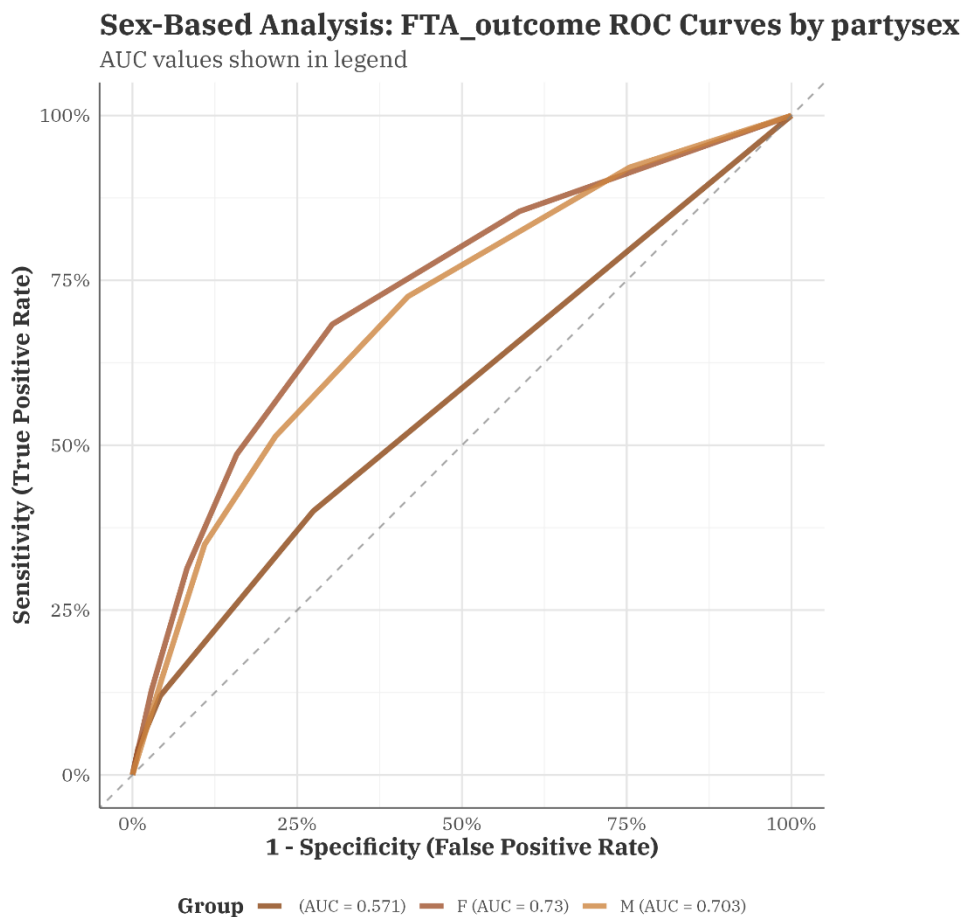
Predictive Validity – Sex-Specific AUC Results

Examining predictive performance across demographic subgroups is needed to identify the potential for algorithmic bias that could accent disparities in different types of outcomes within the criminal justice system. Breaking down the validation by demographic groups can help reveal whether the PSA has similar predictive performance across characteristics such as sex, detect calibration differences that might systematically over- or under-predict risk for specific populations, and provide empirical evidence to determine whether risk scores require population-specific interpretation guidelines or whether universal thresholds maintain both predictive accuracy and fair treatment across diverse defendant populations. Note that we focus on sex as the primary socio-demographic category to be bias-tested given 1) limited sociodemographic reporting data available to us and 2) high rates of missingness observed for race-ethnicity data, which limit capacity to validate the scale, particularly for underrepresented race-ethnicity categories.

Per Figure 7, the sex-based AUC analysis shows relatively consistent predictive performance across gender categories with modest but noteworthy variations that have important implications for equitable risk assessment implementation. For the CAR prediction, the PSA's FTA scale shows good predictive performance for both male and female defendants, with the AUC for female defendants being slightly higher (0.730) compared to males at 0.703, indicating that the PSA may be marginally more effective at predicting court appearance behavior among female defendants. The PSA's NCA scale reveals a reversed

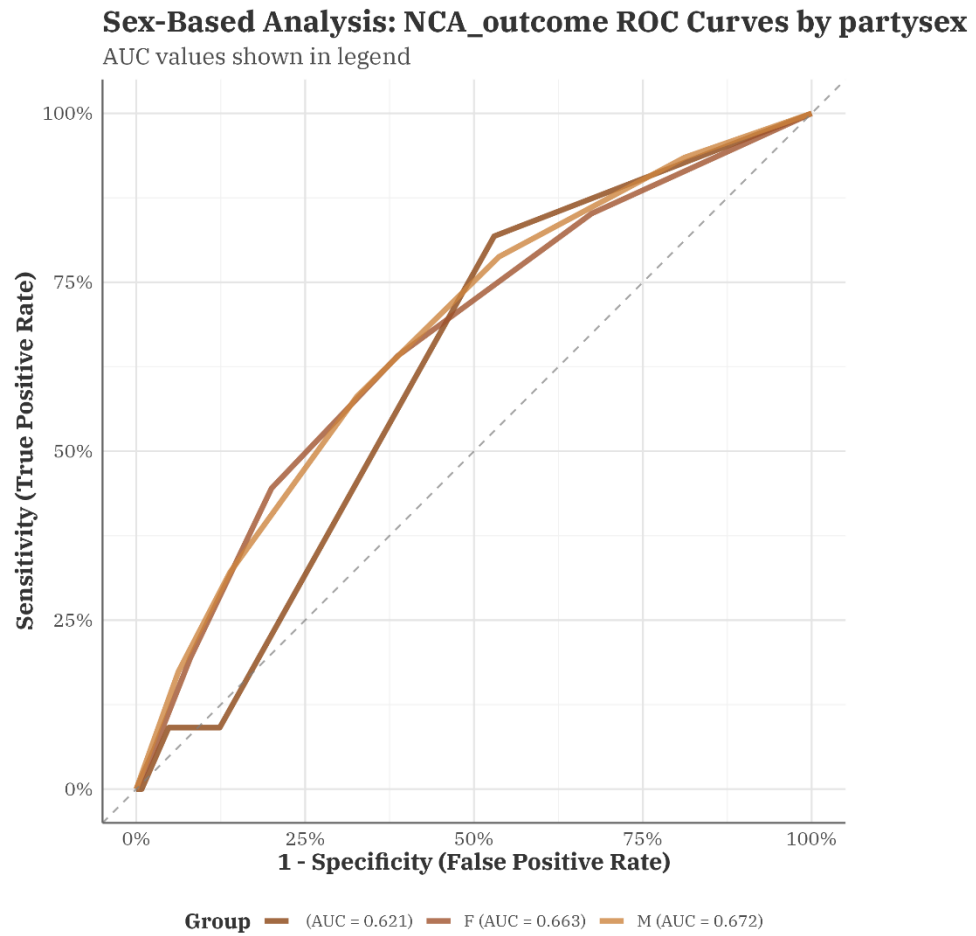
pattern, with the NCA scale demonstrating slightly better predictive performance (AUC = 0.672) compared to female defendants (AUC = 0.663), though both AUCs are within the good range, and the difference is practically minimal. The most pronounced sex-based variation emerges in estimations of the NVR from the NVCA flag, where the flag has limited predictive utility. However, males achieve modestly better discrimination (AUC = 0.588) compared to females (AUC = 0.557), with the latter approaching the threshold for chance-level performance. Together, these findings suggest that while the PSA maintains generally equitable predictive performance across sex categories for court appearance and general criminal activity outcomes, the PSA's capacity to identify violent crime risk may be marginally differentially calibrated by sex.

Figure 11. Sex-Specific ROC Curves for PSA FTA Risk Scores



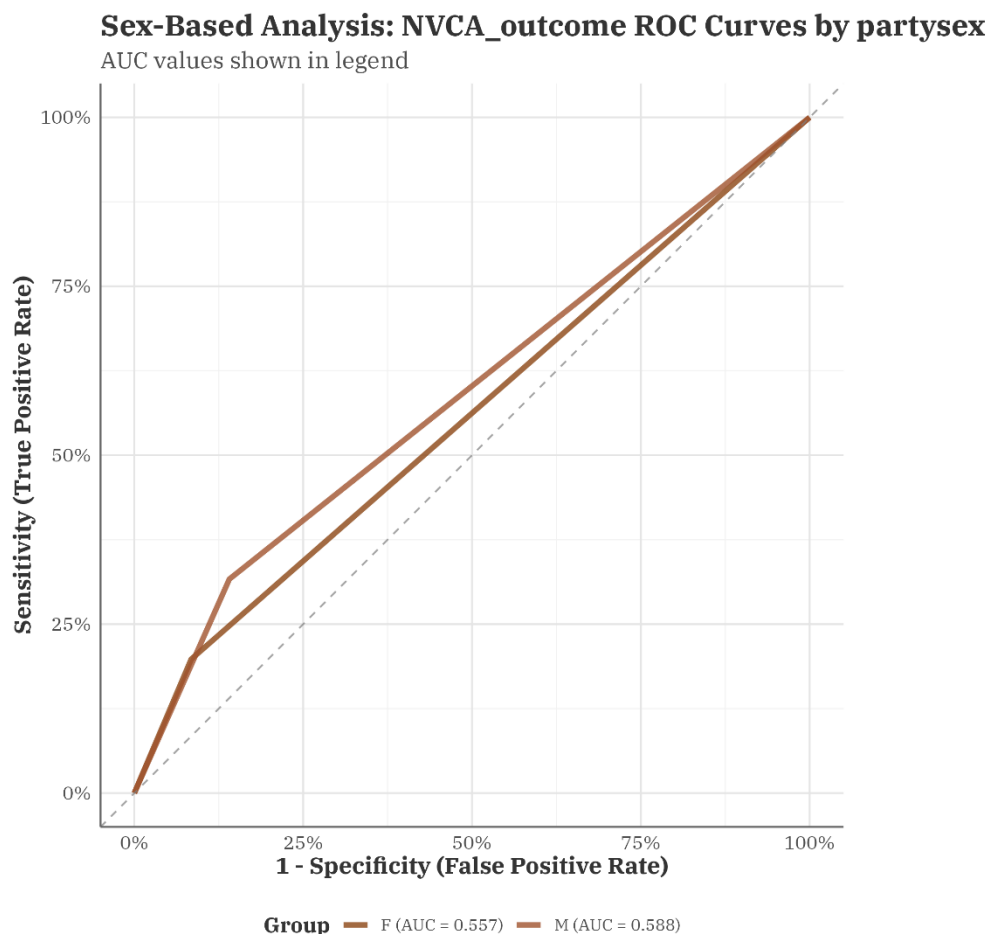
Sex-based ROC curves examine differential FTA risk score performance across gender categories. Analysis reveals potential gender-specific patterns in failure to appear risk prediction supporting equitable pretrial assessment protocols.

Figure 12. Sex-Specific ROC Curves for PSA NCA Risk Scores



Sex-based NCA risk score performance analysis examines gender-specific patterns in criminal activity prediction. Differential performance assessment ensures equitable risk assessment across demographic groups.

Figure 13. Sex-Specific ROC Curves for PSA NCA Risk Scores



Sex-based NVCA risk score analysis evaluates gender differences in violent crime prediction accuracy. Critical assessment for ensuring fair and equitable public safety decision-making across demographic groups.

Predictive Validity - Logistic Regression Results

While descriptive analyses examining outcome rates by PSA score levels and AUC statistics provide important preliminary evidence of predictive validity, these approaches are insufficient for quantifying the independent association between the PSA subscales and pretrial outcomes. Descriptive comparisons fail to account for individual-level, institution-level, and time-related factors that may also influence the relationship between PSA scores and outcomes, potentially leading to biased estimates of the PSA's predictive validity. Most importantly, particularly in relation to the descriptive analysis, the analysis of outcome rates by scores does not disentangle the direct effects of defendant risk characteristics from the indirect effects operating through pretrial supervision decisions, which are themselves influenced by PSA scores but vary systematically across jurisdictions. The hierarchical structure of court systems, where individual courts operate within broader judicial districts with shared administrative practices and policies, creates additional dependencies in the data that violate the independence assumptions of the previous approaches.

To address these methodological complications, we used a regression-based approach using nested random effects models that account for the hierarchical structure inherent in court system data. This modeling strategy recognizes that individual cases are nested within specific courts, which are themselves situated within broader judicial districts, creating multiple levels of potential clustering that could influence both risk score distributions and pretrial outcomes. We incorporated court-specific

random effects nested within district-level random effects to capture this hierarchical variation, while also including year-specific random intercepts to control for temporal trends that might affect both scoring practices and outcome rates over time. This analytical framework helps address a fundamental confounding issue: the apparent relationships between risk scores and pretrial outcomes observed in unadjusted analyses may simply reflect unmeasured characteristics specific to particular courts, districts, or time periods that independently influence both the assignment of risk scores and the likelihood of pretrial failure. By modeling these sources of variation explicitly, we can better isolate the true predictive relationship between risk assessment scores and outcomes, separating genuine predictive validity from spurious associations that arise from shared dependence on unobserved contextual questions.

Second, careful attention to exposure time through log-transformed days between release and case closure addresses the fundamental issue that longer pretrial periods provide greater opportunity for failure events, creating systematic bias if exposure varies by risk level or other characteristics (i.e., if exposure is systematically different across courts or districts, then courts or districts where defendants have higher average exposure times might have lower CAR, PSR, or NVRs). Third, by introducing a control variable for pretrial supervision intensity using the PML (Pretrial Monitoring Level) categories, we were able to adjust for a primary mediating pathway (since it is important to disentangle the associations between PSA scores and pretrial outcomes from a factor that covaries with both scores and outcomes), while district-by-PSA score interaction terms test whether supervision assignment practices vary across jurisdictions in ways that could confound apparent tool effectiveness. Without proper control for this mediation pathway, apparent relationships between PSA scores and outcomes may simply reflect the effectiveness of supervision (that covaries with PSA score level) rather than the underlying predictive validity of the risk assessment tool itself.

The regression analyses reported in Table 11 revealed statistically significant associations between PSA scaled scores and all three pretrial outcome measures, with higher risk scores consistently predicting lower odds of positive outcomes across all domains, even after adjusting for all of these pathways. For each one-unit increase in the FTA scaled score, the odds of court appearance decreased by approximately 25% (OR = 0.753, 95% CI: 0.701-0.809, $p < 0.001$), indicating that defendants with higher FTA risk scores were substantially less likely to make all required court appearances. Similarly, higher NCA scaled scores were associated with reduced odds of public safety, with each one-unit increase corresponding to a 16% decrease in the odds of avoiding new criminal activity during the pretrial period (OR = 0.842, 95% CI: 0.778-0.908, $p < 0.001$). The strongest association was observed for the NVR, where each one-unit increase in the NVCA scaled score was associated with a 64% reduction in the odds of avoiding NVCA (OR = 0.358, 95% CI: 0.249-0.516, $p < 0.001$). These findings suggest that PSA risk scores maintain their predictive validity even after controlling for demographic characteristics, pretrial supervision intensity, exposure time, and the nested structure of court systems, providing preliminary evidence for the PSA's relative effectiveness in identifying defendants at varying levels of pretrial risk.

Table 11. Adjusted Odds Ratios from Logistic Regressions for PSA Outcomes

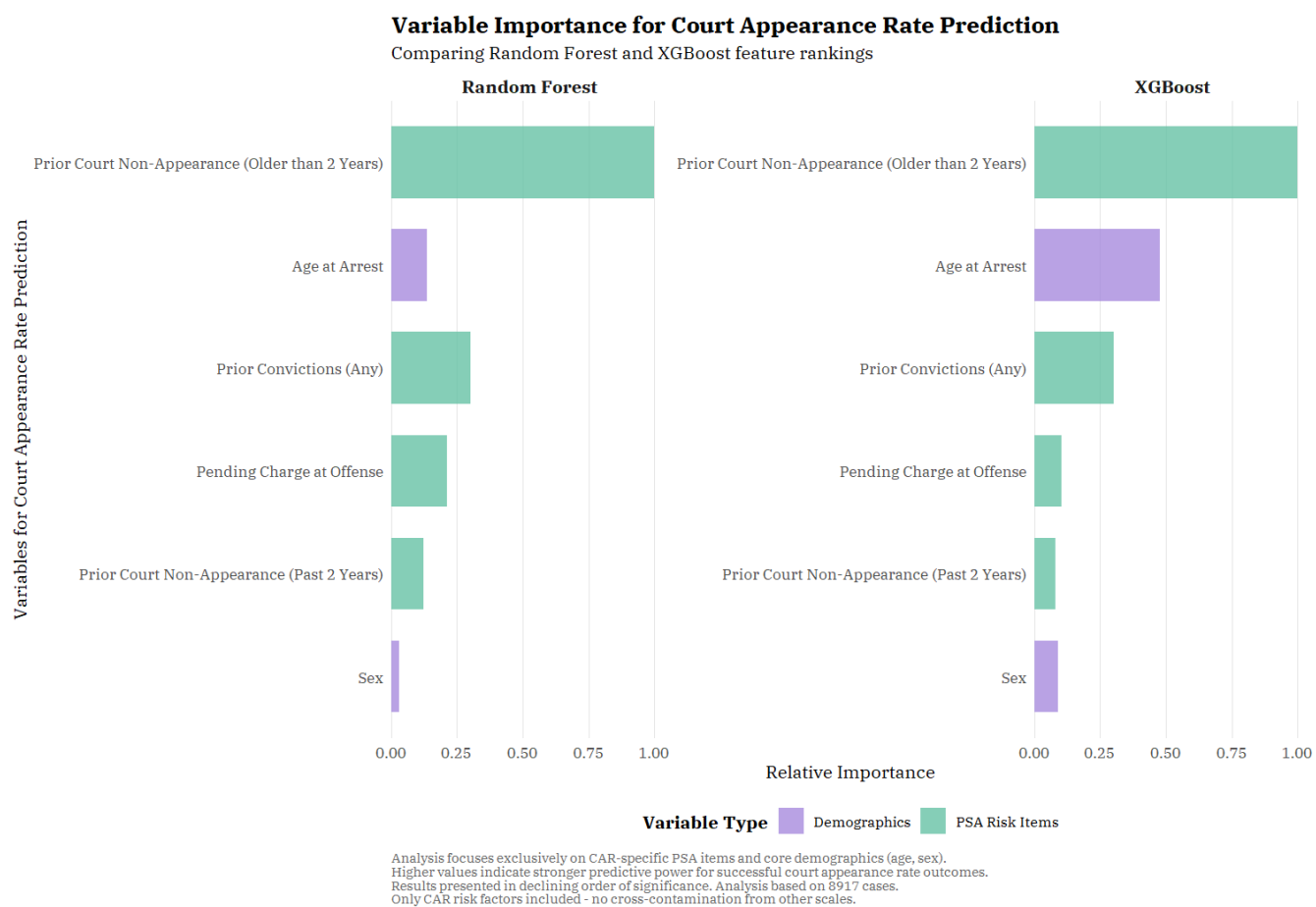
Outcome Measure	Odds Ratio (95% CI)*	P-Value
Court Appearance Rate (CAR)	0.753 (0.701-0.809)	<0.001
Community Safety Rate (PSR)	0.842 (0.78-0.908)	<0.001
Non-Violence Rate (NVR)	0.358 (0.249-0.516)	<0.001

Which PSA Factors Best Predict Pretrial Outcomes

Finally, we conducted variable importance analyses using machine learning approaches to identify which scored factors on each PSA subscale were most strongly predictive of pretrial outcomes. For each outcome, we restricted our models to include only the outcome-specific PSA risk factors and core demographic variables that fall outside the PSA, such as age at arrest and sex. We applied both Random Forest and XGBoost machine learning algorithms to quantify the relative contribution of each predictor variable to model performance. Variable importance scores, which measure how much each predictor contributes to reducing prediction error, were normalized within each model to facilitate direct comparison between the two algorithms. We present these results on a standardized 0-1 scale where higher values indicate greater relative importance in predicting the outcome of interest. We use two different approaches to increase confidence in the identification of the most influential predictive factors.

Our variable importance analysis for CAR prediction revealed consistent patterns across both Random Forest and XGBoost models, with historical court appearance emerging as the strongest predictor. We found that Prior Court Non-Appearance (Older than 2 Years) had the highest relative importance (1.0) in both algorithms, indicating that past court attendance patterns remain highly predictive of future appearance regardless of temporal distance. Age at arrest showed moderate importance as a demographic factor, while Prior Convictions (Any) and Pending Charge at Offense displayed similar moderate predictive power. We observed that Prior Court Non-Appearance (Past 2 Years) and Sex exhibited the lowest importance scores, suggesting that relative to other items on the subscale, recent non-appearance history and sex contribute minimally to court appearance success prediction.

Figure 14. Variable Importance Plot (CAR)

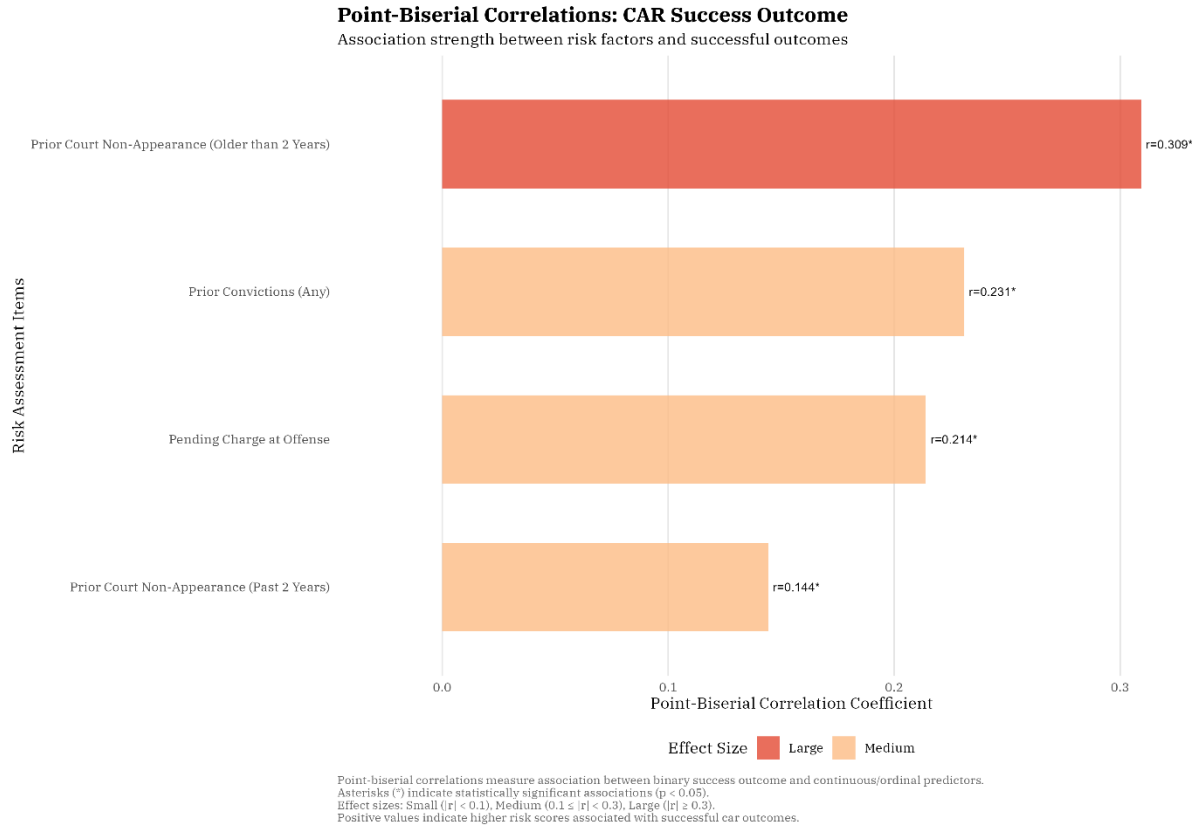


Correlation Matrix: CAR Success Factors

Pearson correlations between risk assessment items and demographics

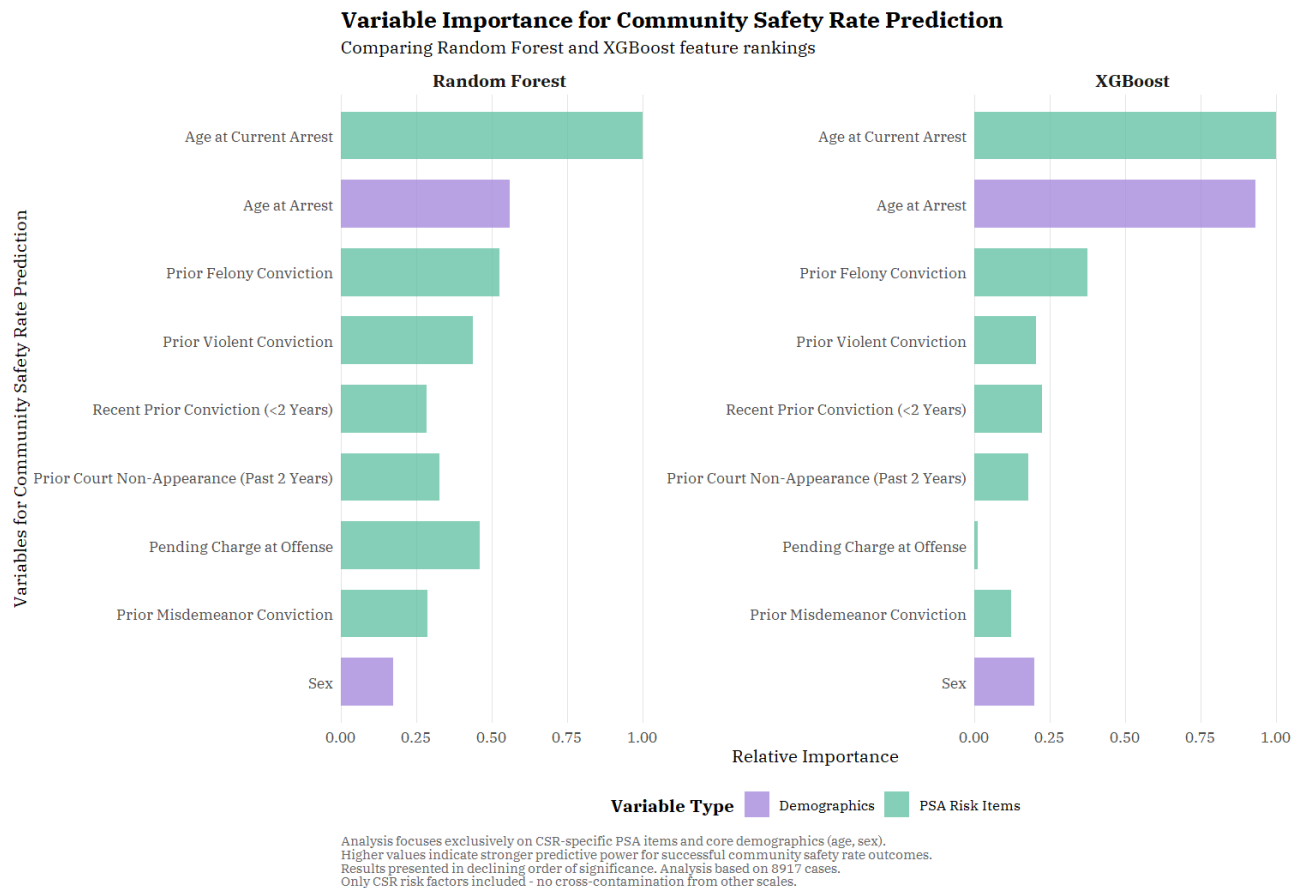


Analysis includes CAR-specific PSA items, success outcome variable, and demographic factors. Values represent Pearson correlation coefficients ranging from -1 (perfect negative) to +1 (perfect positive). Stronger correlations indicate greater linear relationships between variables. White cells indicate weak correlations, colored cells indicate moderate to strong associations.



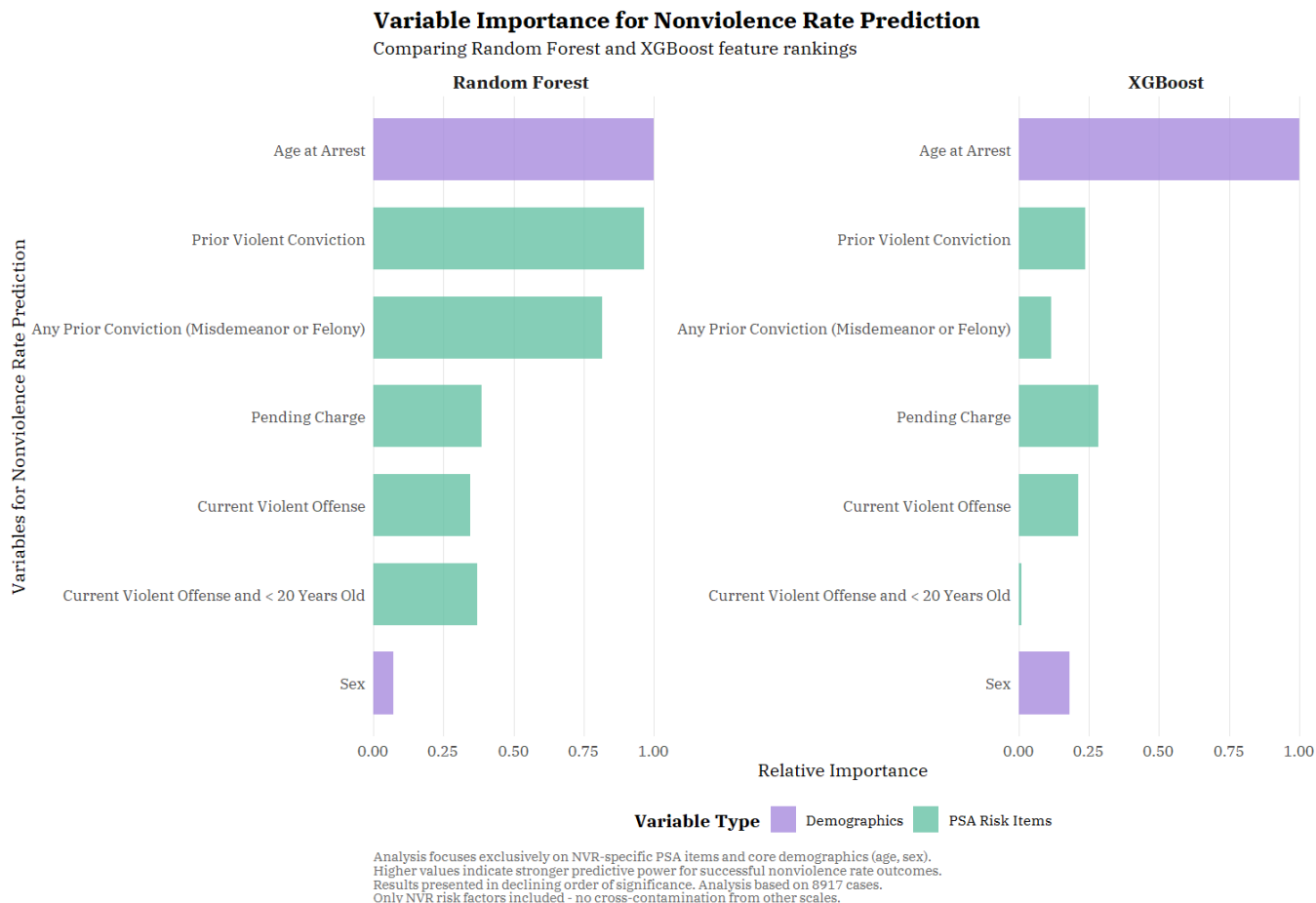
Our PSR variable importance plots suggest that age-related factors were the most important predictors, with the Age at Current Arrest variable achieving maximum relative importance (1.0) across both machine learning models. Prior Felony Conviction was the second most important predictor, followed by Prior Violent Conviction and Recent Prior Conviction, indicating that criminal history severity and recency significantly influence PSR outcomes. We observed mixed importance for demographic factors, with Age at Arrest displaying moderate predictive power while Sex contributed minimally to model performance.

Figure 15. Variable Importance Plot (PSR)



Finally, the NVCA flag for NVR prediction had the most balanced importance distribution across predictors, with Age at Arrest as the primary demographic predictor achieving maximum relative importance (1.0) in both models. We found that Prior Violent Conviction demonstrated high importance, reinforcing the predictive validity of specialized violence-related risk factors for this outcome domain. Any Prior Conviction (Misdemeanor or Felony) and violence-specific items (Current Violent Offense, Pending Charge) showed moderate importance levels, suggesting that both general criminal history and violence-specific factors contribute meaningfully to nonviolence predictions. We observed that the XGBoost model showed greater selectivity compared to Random Forest, with several variables displaying substantially reduced importance scores, indicating potential algorithmic differences in how violence prediction models weight different risk factors. We consistently found that Sex showed minimal predictive value across both algorithms, suggesting limited sex-based differences in nonviolence rate prediction within our sample.

Figure 16. Variable Importance Plot (NVR)



Discussion

In this report, we present preliminary evidence that summarizes the predictive validity of the PSA across New Mexico's Eleventh and First judicial districts. Our examination of the relationship between PSA risk scores and observed pretrial behavior revealed predictive gradients, with successful court appearance rates declining from 90.0% among defendants classified as being at the lowest risk of FTA to 48.5% for those classified as being at the highest risk. Similarly, we observed PSRs declining from 92.0% for NCA Score 1 defendants to 59.6% for those with NCA Score 6. These findings suggest that the PSA successfully identified meaningful risk stratification across the PSA's scoring range, providing initial support for the PSA's core theoretical structure.

We also generated ROC analyses that produced AUC statistics of 0.713 for court appearance, 0.678 for community safety prediction, and 0.584 for violence-free outcomes. Our multivariate logistic regression analyses, which included statistical adjustments for defendant demographics, supervision intensity, exposure duration, and court jurisdiction clustering, indicated statistically significant associations between PSA scores and all three outcome measures.

However, several limitations affect the interpretation and generalizability of these findings but also suggest frontiers for future research. We were missing race-ethnicity data for 64.6% of the final sample, preventing meaningful differential validity analyses by race-ethnicity. Additionally, we had to limit our primary predictive validity analysis to defendants from two of the six judicial districts implementing the PSA due to unreliable jail booking and release log data, which reduced the sample from 53,268 to 10,872 cases and potentially limited generalizability to jurisdictions with different characteristics. The geographic

concentration of cases, with 75.9% originating from the Thirteenth Judicial District, further constrains a pure assessment of cross-jurisdictional variation in PSA performance at this point.

The evidence presented here tentatively supports the continued use of the PSA as a decision-support tool for pretrial release and supervision decisions across New Mexico's judicial districts yet also highlights areas for improvement and the need for better data measurement and reporting processes to increase confidence in the generalizability of the present findings across PSA-implementing jurisdictions in New Mexico. Strong predictive validity for court appearance outcomes (AUC = 0.713) and acceptable predictive validity for general criminal activity prediction (AUC = 0.678) suggest that the PSA may be a helpful, directionally accurate, discretionary tool for informing evidence-based pretrial release decision-making. However, the more limited predictive validity of the PSA when it comes to predicting violent crime outcomes pretrial (AUC = 0.584), a finding which replicates across other validation studies (Severson & Ferguson, 2024), suggests that additional risk factors or assessment approaches may be needed for public safety decisions involving potential violent recidivism, potentially coupled with a need to quantify and communicate the greater scope of uncertainty in the NVCA flag's predictive validity to relevant pretrial parties. We encourage jails in the state of New Mexico in PSA-implementing districts to use standardized, comprehensive data collection systems for their booking and demographic data to ensure accuracy of state-wide results, and the systematic and regular collection of complete demographic data, such that future validation and revalidation efforts of the PSA, if undertaken in New Mexico, can more systematically test for predictive bias across other socio-demographic factors as a fairness check on the tool.

Finally, we want to be explicit about the inherent role of discretion and risk tolerance in pretrial decision-making contexts, recognizing that no PRAI or human judgment achieves perfect predictive accuracy. Even when PRAIs attain reasonable levels of predictive validity, as we report with the PSA's performance for court appearance and general public safety, they will inevitably produce both false-positive and false-negative predictions that result in seemingly incorrect decisions when viewed retrospectively. We argue that this limitation reflects a fundamental challenge of predicting human behavior in complex social contexts rather than a specific deficiency of the PSA per se, and human decision-makers operating without actuarial tools face identical predictive constraints while potentially exhibiting additional biases related to cognitive heuristics, personal experience, or demographic characteristics that may systematically influence risk judgments in ways that do not align with empirical evidence.

In our view, the question for the AOC and court systems more generally is not whether prediction errors will occur, but rather, how to minimize harmful errors while maintaining appropriate levels of release that respect due process principles and avoid unnecessary detention. We suggest that stakeholders explicitly consider their tolerance for different types of prediction errors, recognizing that overly restrictive approaches may lead to excessive pretrial detention (which previous studies, conducted by CARA, suggest is not cost-effective). In contrast, overly permissive approaches may lead to lower rates of pretrial success and a decrease in public safety. We argue that the PSA is best viewed as one component of a broader decision-making framework that combines actuarial information with judicial discretion and case-specific considerations, rather than as a deterministic tool designed to perfectly predict pretrial outcomes, allowing decision-makers to have baseline risk information at least while retaining flexibility to incorporate case-specific factors and professional judgment that may not be captured in the PSA. However, we also want to note that there are domains in which the PSA's predictive performance and/or how PSA results are communicated to stakeholders could be improved (e.g., provide scores and predicted probabilities of FTA, NCA, and NVCA outcomes to stakeholders; communicate the degree of uncertainty surrounding estimates of FTA, NCA, and NVCA) (see Moore et al., 2025).

The statewide expansion of the PSA across New Mexico's diverse jurisdictional contexts provides a potential model for other states, while we believe that the identified limitations offer important lessons for implementation elsewhere. As New Mexico expands PSA implementation and refines data collection procedures over time, regular revalidation studies (i.e., every two years) will help to ensure continued effective and equitable performance across the state's evolving and expanding pretrial population.

References

- AdvancingPretrial.org. (2020). About the Public Safety Assessment (PSA). Retrieved from: <https://advancingpretrial.org/psa/about/>
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(1), 8753-8830.
- Bornstein, B. H., Tomkins, A. J., Neeley, E. M., Herian, M. N., & Hamm, J. A. (2013). Reducing courts' failure-to-appear rate: A procedural justice approach. *Journal of Empirical Legal Studies*, 10(4), 777-806.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cohen, T. H., & Reaves, B. A. (2007). Pretrial release of felony defendants in state courts (NCJ 214994). Bureau of Justice Statistics.
- Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*, 13(3), 206-222.
- DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., & Misra, S. (2018). *The Public Safety Assessment: Predictive utility and differential prediction by race and gender in Kentucky*. RTI International.
- Desmarais, S. L., & Singh, J. P. (2013). Risk assessment instruments validated and implemented in correctional settings in the United States: An empirical guide. *Behavioral Sciences & the Law*, 31(6), 708-721.
- Ferguson, E., de la Cerda, H., Guerin, P., & Moore, C. (2021). *Bernalillo County public safety assessment validation*. Institute for Social Research, Center for Applied Research and Analysis, University of New Mexico.
- Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation*, 80(2), 38-46.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Greiner, D. J., Stubenberg, M., & Halen, R. (2020). *Validation of the PSA in Harris County, TX*. Harvard Law School.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.
- Laura & John Arnold Foundation. (2016). *Developing a national model for pretrial risk assessment*. Arnold Ventures.
- Lowenkamp, C. T., Lemke, R., & Latessa, E. (2013). *The development and validation of a pretrial screening tool* (Research report). Arnold Ventures.
- Lowenkamp, C. T., VanNostrand, M., & Holsinger, A. (2013). *The hidden costs of pretrial detention*. Arnold Ventures.
- National Association of Pretrial Services Agencies. (2020). *Standards on Pretrial Release: Revised 2020*. NAPSA.

Severson, A., & Ferguson, E. (2024). *A revalidation study of Bernalillo County's Public Safety Assessment*. Institute for Social Research, Center for Applied Research and Analysis, University of New Mexico.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4), 680-712.

Skog, A., & Lacoë, J. (2021). *Validation of the PSA in San Francisco*. eScholarship, University of California.

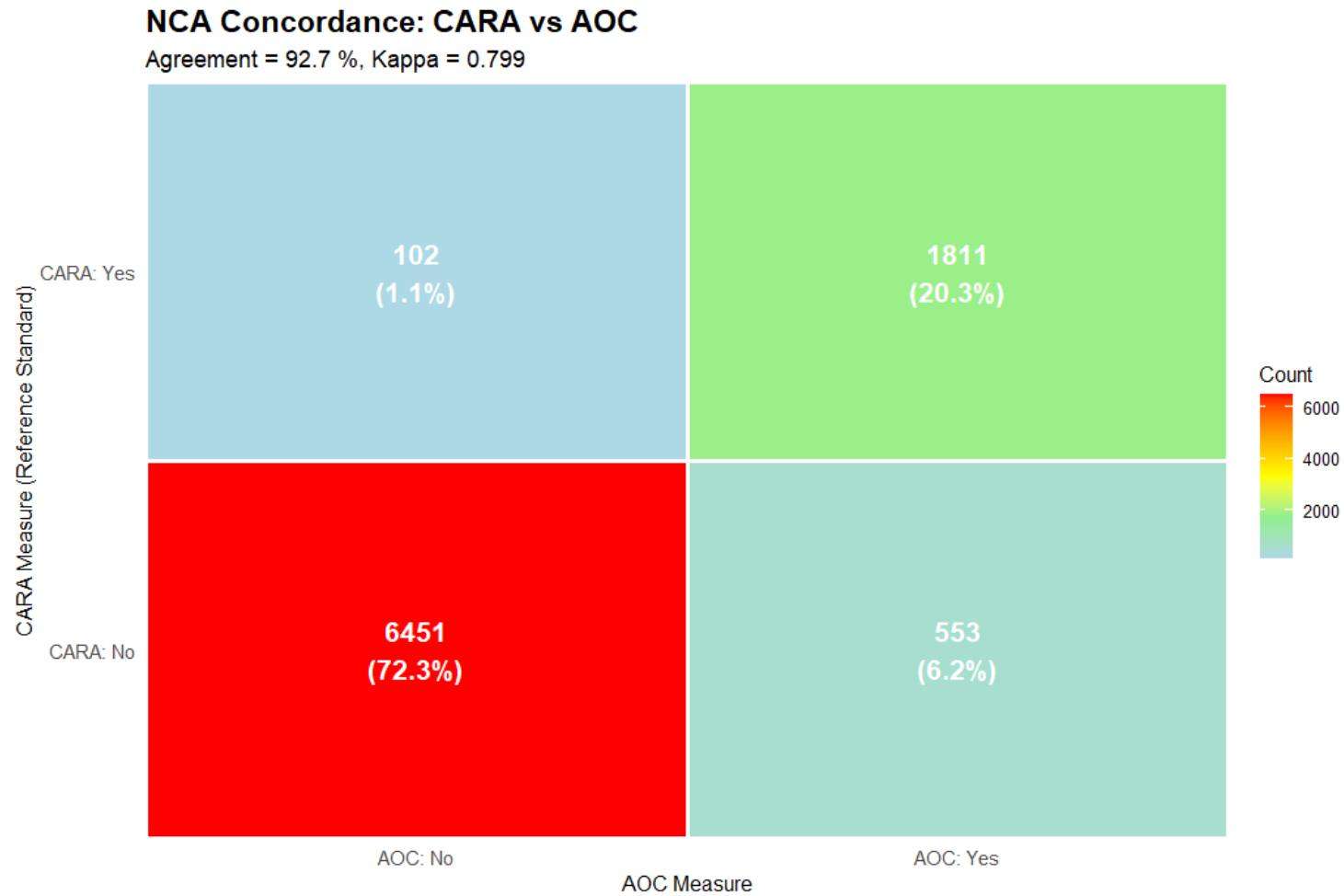
Appendix A – Scope of Concordance Between AOC’s Reported Outcome Measures with CARA’s for the 1st and 13th Judicial Districts

We evaluated whether the pretrial outcomes recorded in the AOC database matched those we found when we independently reviewed court records through manual lookups, specifically focusing on jurisdictions where we had higher confidence in the quality of the jail booking and release data. This type of testing (i.e., concordance testing) is an important robustness check because any study that aims to explore how well the PSA predicts pretrial outcomes depends on having confidence in the reliability of the measurement of outcomes. If administrative data misclassifies pretrial outcomes, we could reach misleading conclusions about the PSA’s predictive ability if we relied solely on administrative data to drive the analysis.

To this end, we explored concordance (i.e., the degree to which AOC’s classification and CARA’s manual verification produced the same classifications) across all three pretrial outcomes: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA). We calculated how often the two approaches (i.e., AOC’s classifications and our manual-lookup-generated calculations) aligned on individual cases, used kappa statistics to determine whether this agreement was meaningful since even random classification would occasionally match by chance alone, and assessed correlation coefficients to understand whether the two measurement approaches produced similar patterns across all cases.

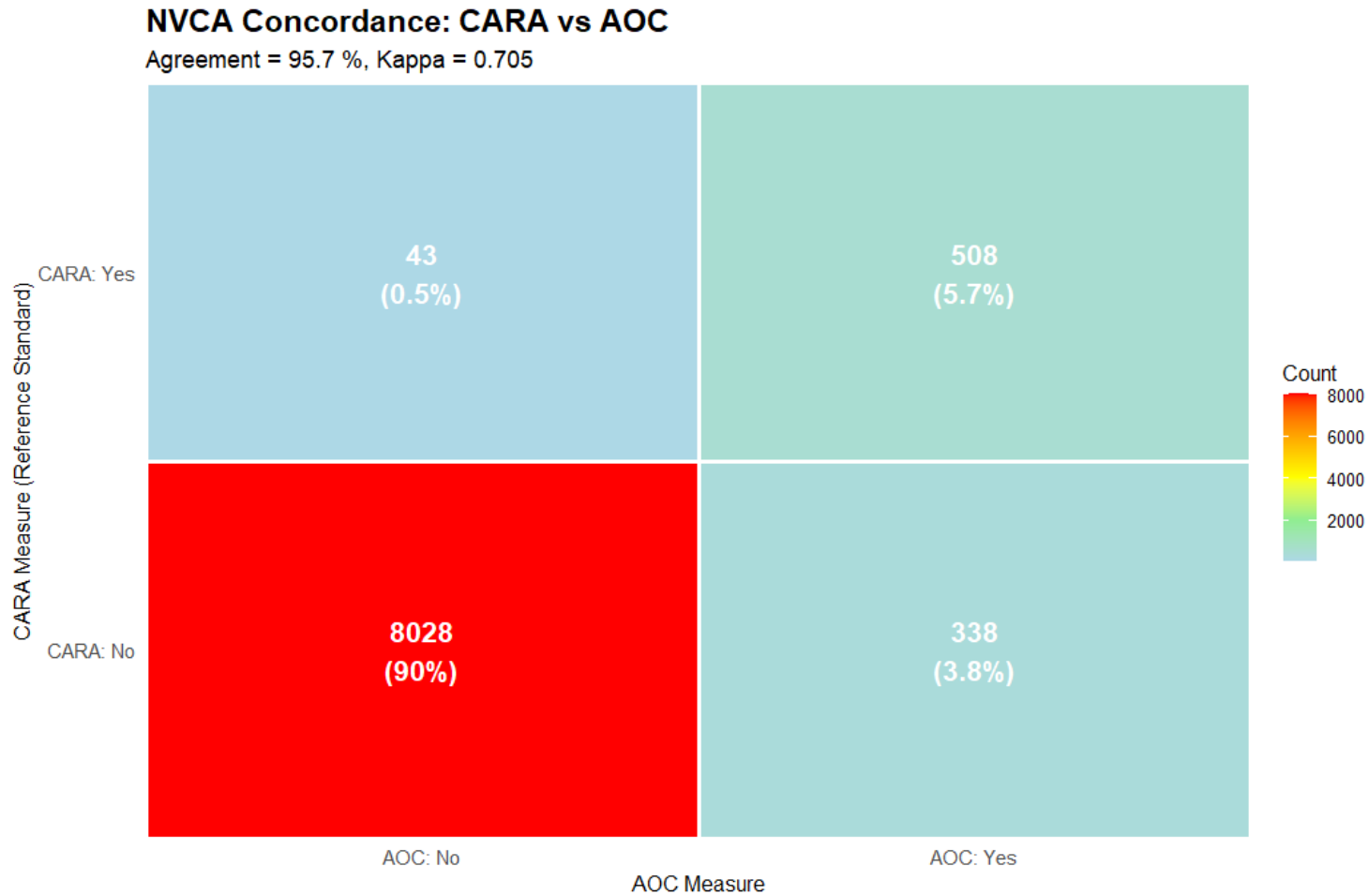
We found that the records provided by AOC and our own lookups matched in 99% of FTA cases, 92.7% of NCA cases, and 95.7% of NVCA cases. The kappa statistics ($\kappa = 0.799$ for NCA and $\kappa = 0.705$ for NVCA) confirmed that this agreement exceeded what we would expect by chance alone. Kappa values range from 0 to 1, where 0 indicates no agreement beyond random chance and 1 indicates perfect agreement, with values above 0.60 generally considered substantial and values above 0.80 considered near-perfect. Our kappa values of 0.799 and 0.705 fall in the substantial to near-perfect range, meaning the two measurement approaches produced highly consistent classifications that we can trust to reflect real agreement rather than coincidence. Our correlation analysis showed a strong positive relationship ($r = 0.807$) between the two measurement approaches for NCA outcomes, meaning that when one approach identified more criminal activity, the other typically did as well. When we examined agreement patterns separately for each judicial district, we found that some jurisdictions showed closer alignment between the administrative records and manual verification than others. These findings suggest that while the two measurement approaches generally produced consistent results, differences in how individual districts record and process case outcomes may affect the reliability of the data used for PSA validation, underscoring the need for standardized outcome measurement practices and ongoing data quality monitoring in pretrial risk assessment systems.

Figure x. NCA Concordance Scatter Plot



Confusion matrix visualization shows agreement between CARA validated new criminal activity outcomes and AOC administrative charge-free status data. Diagonal cells show cases where both measurement approaches agree on NCA classification. Off-diagonal cells highlight discrepancies that may reflect differences in data collection methods, coding procedures, or outcome verification processes. Understanding concordance patterns supports data quality assessment and informs decisions about optimal outcome measurement approaches for pretrial criminal justice research and evidence-based policy development.

Figure x. NVCA Concordance Scatter Plot



NVCA concordance analysis between CARA validated violent crime outcomes and AOC new violent charge administrative data reveals agreement patterns for high-stakes public safety decisions. Perfect agreement on diagonal shows consistent identification of violent criminal activity across measurement approaches. Disagreement cells identify cases requiring additional investigation. Understanding measurement concordance for violent crime outcomes supports confident implementation of detention and supervision decisions that prioritize community safety while ensuring accurate risk assessment based on reliable outcome measurement.

Figure x. Observed FTA Rate by AOC Data versus CARA Manual Pulls

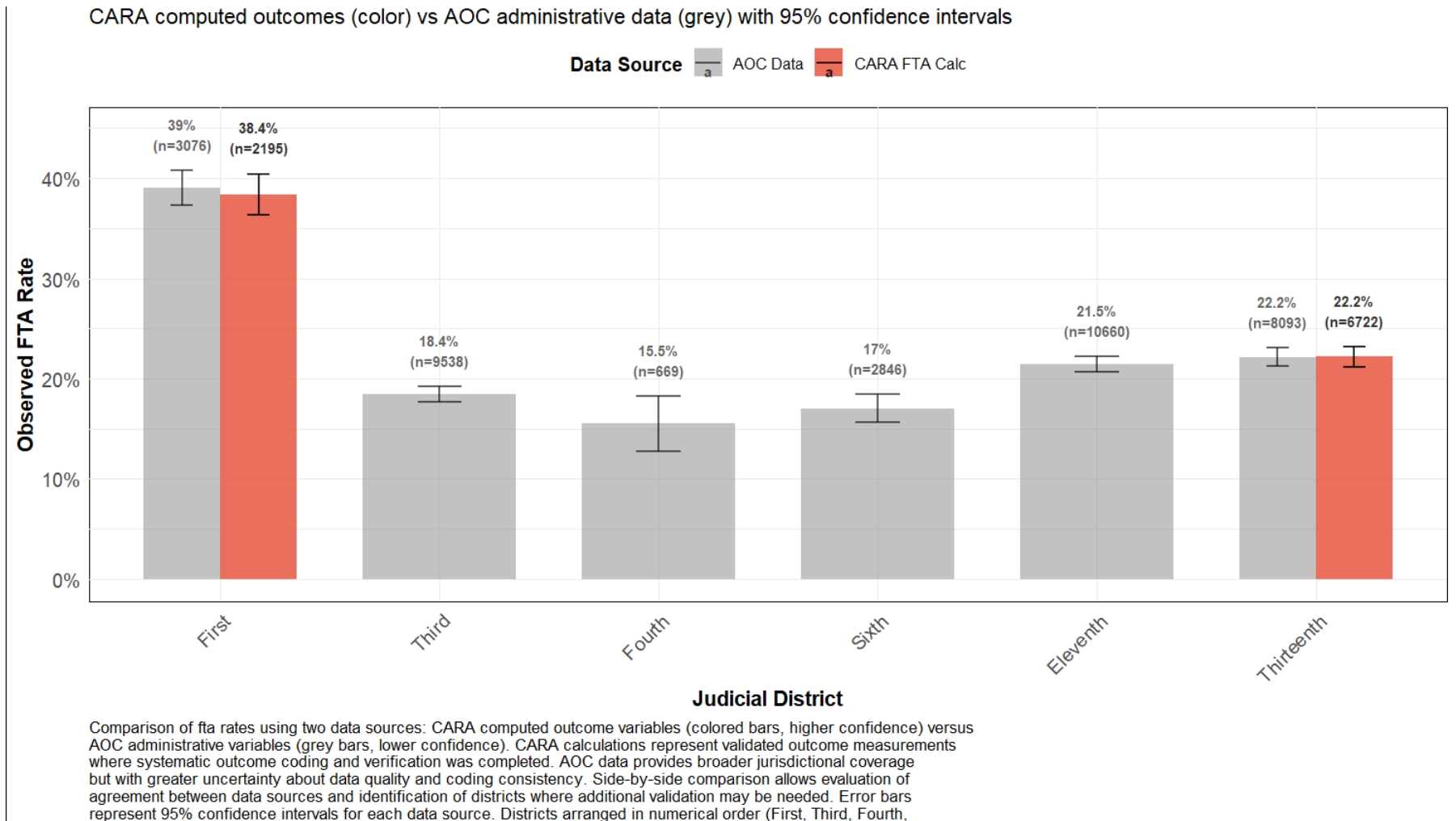
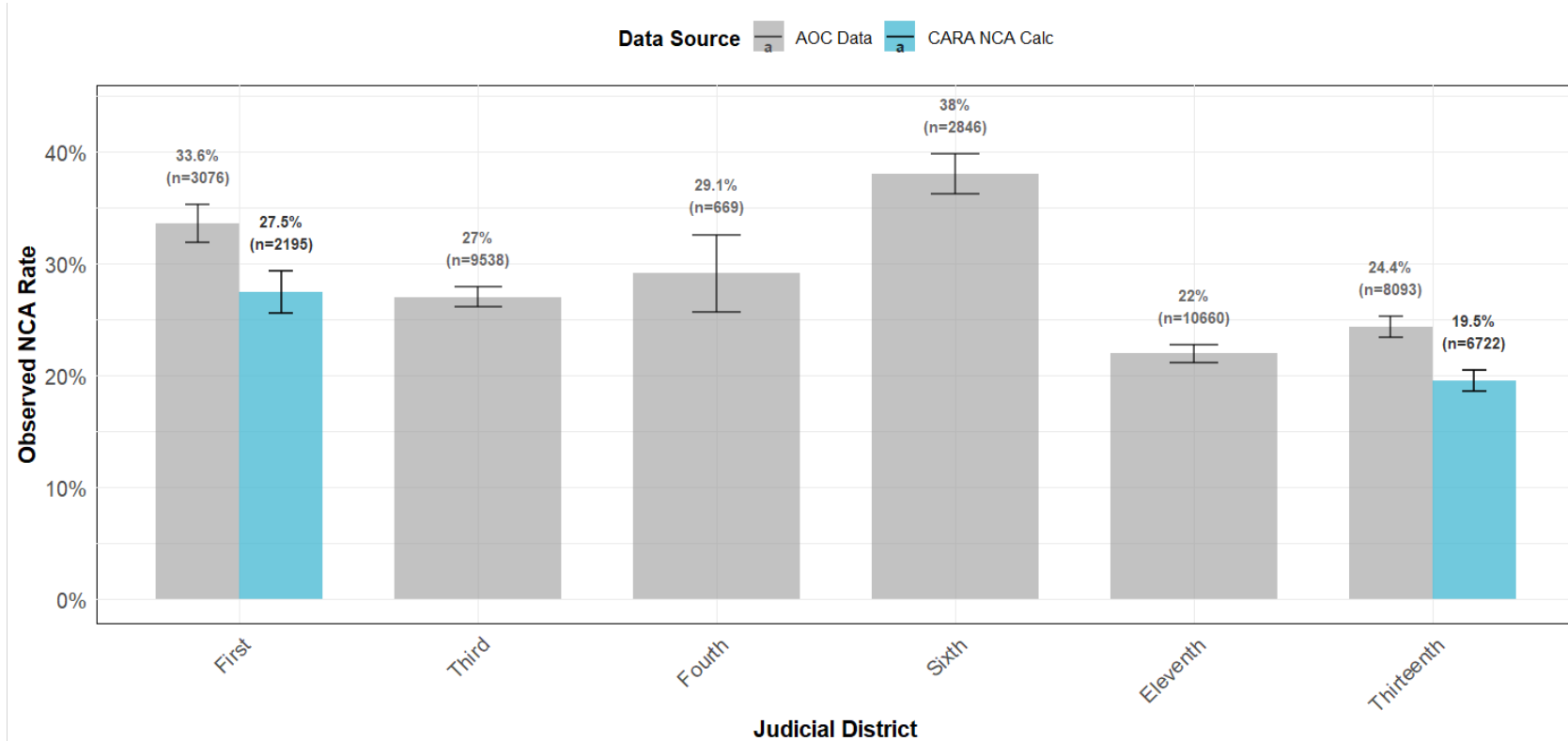
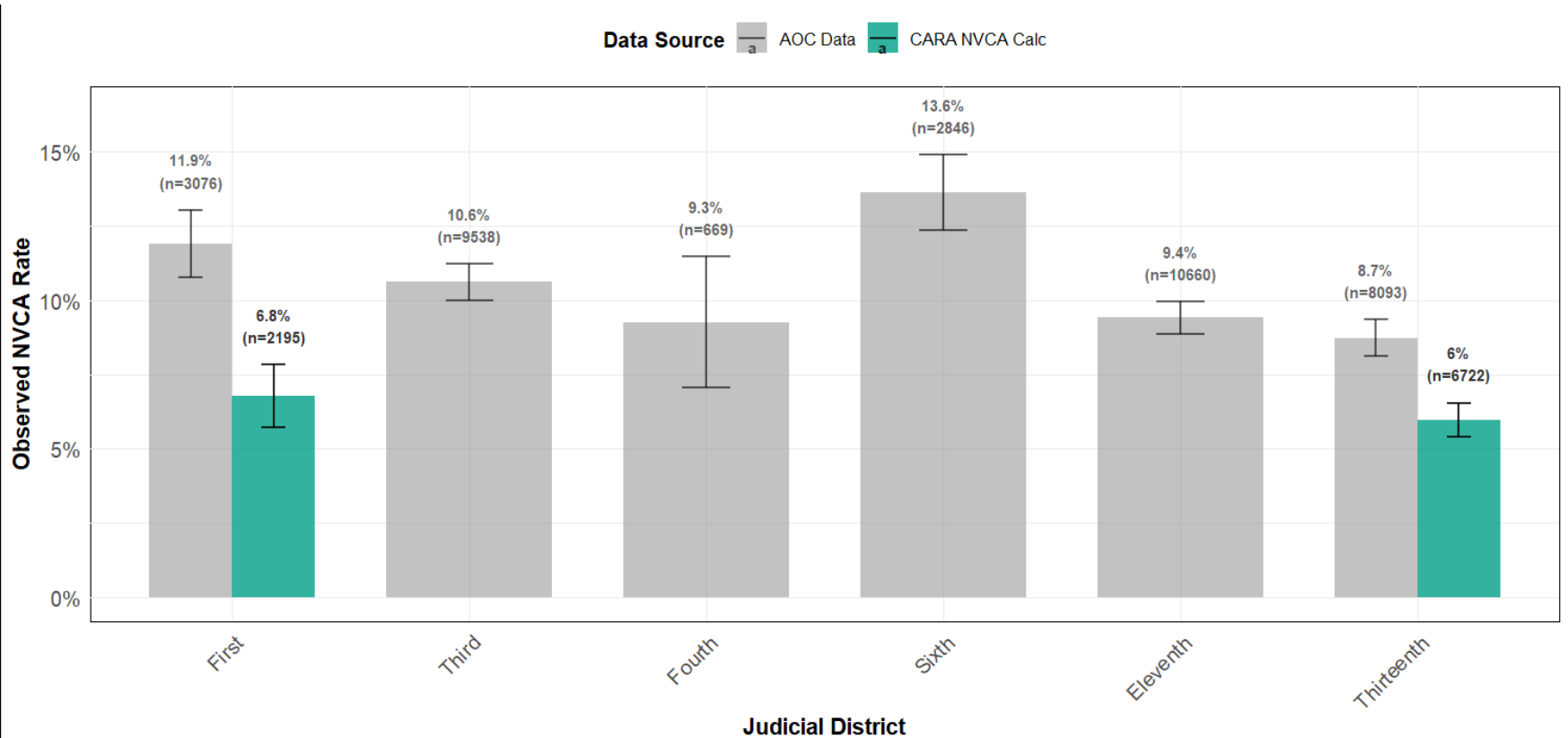


Figure x. Observed NCA Rate by AOC Data versus CARA Manual Pulls



Comparison of nca rates using two data sources: CARA computed outcome variables (colored bars, higher confidence) versus AOC administrative variables (grey bars, lower confidence). CARA calculations represent validated outcome measurements where systematic outcome coding and verification was completed. AOC data provides broader jurisdictional coverage but with greater uncertainty about data quality and coding consistency. Side-by-side comparison allows evaluation of agreement between data sources and identification of districts where additional validation may be needed. Error bars represent 95% confidence intervals for each data source. Districts arranged in numerical order (First, Third, Fourth, etc.) for consistent presentation across analyses.

Figure x. Observed NVCA Rate by AOC Data versus CARA Manual Pulls



Comparison of nvca rates using two data sources: CARA computed outcome variables (colored bars, higher confidence) versus AOC administrative variables (grey bars, lower confidence). CARA calculations represent validated outcome measurements where systematic outcome coding and verification was completed. AOC data provides broader jurisdictional coverage but with greater uncertainty about data quality and coding consistency. Side-by-side comparison allows evaluation of agreement between data sources and identification of districts where additional validation may be needed. Error bars represent 95% confidence intervals for each data source. Districts arranged in numerical order (First, Third, Fourth, etc.) for consistent presentation across analyses.