

**Validation of the COMPAS at
the Bernalillo County
Metropolitan Detention
Center: A Four-Year
Retrospective Analysis of
Predictive Accuracy and
Reliability**

Prepared By:

Alex Severson, Ph.D.

Prepared For:

The Bernalillo County Metropolitan
Detention Center (MDC)

January 2026

Introduction

Risk assessment instruments (RAIs) have been increasingly used in correctional systems to estimate the probability of various outcomes, including recidivism, treatment program success, and institutional misconduct. The increased adoption of RAIs over the past 20 years represents a move away from unstructured, clinical judgment in evaluating a given individual's risk level toward a more probability-based, statistically informed evaluation of risk (Andrews & Bonta, 2010; Austin et al., 2013). This growth reflects increased recognition that data-driven RAIs can enhance institutional safety and improve institutional resource allocation (Hamilton, Kigerl, & Hummer, 2021; Smeekens et al., 2024; Russo et al., 2020). Recent research indicates that RAIs, on balance, generate more accurate and less biased predictions than unstructured professional judgment, as they help reduce evaluators' subjective biases and generally provide consistent, data-driven risk estimates (Viljoen et al., 2025; Brookings Institution, 2023).

Despite the potential benefits of RAIs, implementing any given RAI requires ongoing validation to ensure its predictive accuracy remains robust across different populations, time periods, and institutional contexts. External validation studies suggest that RAIs are often less predictive in new settings than during initial validation (Fazel et al., 2022). Additionally, many commercial RAIs operate as black-box systems, where the use of proprietary algorithms obscures the features and weightings used to calculate risk scores, limiting third-party observers' ability to scrutinize decisions and preventing systematic assessment of potential biases (Brookings Institution, 2023; Skeem & Lowenkamp, 2025). For these reasons, continuous monitoring of a given RAI's predictive performance is necessary to detect any degradation in model performance over time and across populations and to ensure that risk classifications remain calibrated to actual rates of institutional misconduct (Applegarth et al., 2023; Bureau of Justice Assistance, 2020).

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) RAI is one of the most widely adopted RAIs in correctional environments, used across hundreds of jurisdictions nationwide for pretrial, sentencing, and correctional management decisions (Brennan et al., 2009; Equivant, 2019). Initially developed by Northpointe Inc. (now Equivant), COMPAS combines information from 137 questions to produce risk classifications ranging from low-risk to high-risk across various outcome measures. The COMPAS includes a specific institutional misconduct scale designed to predict misconduct in jail and prison facilities, though validation studies of the COMPAS have predominantly focused on evaluating its ability to predict post-release recidivism rather than institutional misconduct (Brennan et al., 2009; Long, 2020).

The COMPAS collects information across multiple domains, including criminal history, social environment, personality traits, and attitudes. COMPAS Core comprises 43 scales organized into risk scales (designed for prediction) and need scales (designed to identify intervention targets), with different risk models drawing on distinct subsets of items relevant to their respective outcomes (Northpointe Inc., 2015). For example, the Violent Recidivism Risk Scale relies on five inputs: history of violence, history of noncompliance, vocational/education problems, current age, and age at first arrest (Brennan et al., 2009). For each scale, COMPAS generates raw scores that are converted to decile scores ranging from 1 to 10 based on normative distributions, with scores of 1-4 typically classified as low risk, 5-7 as medium risk, and 8-10 as high risk, though threshold cutoff points can vary across jurisdictions (Dieterich et al., 2016). While Northpointe has published the general structure of its scoring algorithms, including the variables that enter each risk scale, the specific statistical weights applied to individual factors remain proprietary, limiting the ability of external researchers to fully replicate how item-level responses contribute to final risk classifications (Rudin et al., 2020).

The ability to predict institutional conduct within correctional facilities is important because disruptive behavior affects facility safety, staff resources, and operational efficiency, and may influence offender rehabilitation outcomes. Research consistently indicates that institutional misconduct varies significantly

across offender populations, with factors including age, criminal history, antisocial personality disorder, and substance use disorders serving as key predictors of misconduct (Gendreau et al., 1997; Steiner & Wooldredge, 2015). Accurate identification of high-risk individuals can help correctional administrators implement risk- and needs-targeted programming, adjust housing assignments, and allocate security resources more effectively, potentially reducing both misconduct incidents and associated costs. Because institutional environments and offender populations change over time, continual revalidation of RAIs is necessary to ensure their ongoing predictive utility and to identify potential changes in risk factor relationships.

The Bernalillo County Metropolitan Detention Center (MDC) implemented the COMPAS tool, with the most recent validation conducted in 2020 (Jackson & Mendoza, 2020). Since this initial validation, the MDC has experienced several changes in average daily population (ADP), operational procedures, and external factors that could, in theory, influence the potential association between COMPAS scores and institutional misconduct. Specifically, these changes include modifications to pretrial detention policies, an evolving population (e.g., the proportion of jail bookings for felonies has increased since the 2020 Northpointe validation study was completed), and the series of effects of the COVID-19 pandemic on correctional operations. The extent of these broader changes to the institutional and policy environments, coupled with the generally recommended timeline for risk-assessment revalidation, suggests the need for an updated analysis of COMPAS' predictive performance within the MDC.

We address this need by conducting a revalidation study of the COMPAS within the Bernalillo County MDC, using data on bookings and classifications from April 1, 2021, to December 31, 2024, specifically examining how well the COMPAS predicted institutional misconduct. We examined the association between COMPAS scores and institutional misconduct after adjusting for factors that may influence institutional behavior outcomes (e.g., length of stay), and we stratify the sample by age, and sex to see whether the associations vary conditional on demographic characteristics. Finally, we examined patterns of misconduct concentration by empirically testing whether institutional misconduct follows the Pareto principle, with a small subset of inmates accounting for a disproportionate share of misconduct incidents. Understanding the degree of misconduct concentration helps contextualize the practical usefulness of risk classification, as effective identification of high-frequency offenders has direct implications for housing assignments and resource allocation.

Overview of MDC Classification Policy

The Bernalillo County Metropolitan Detention Center's classification policy (ICL 17.00) establishes a framework for assigning inmates to one of nine custody levels (Level 1/High through Level 9/Very Low Minimum) within three security categories (Maximum, Medium, and Minimum). The policy mandates that classification decisions be based on "objective and identifiable criteria" using a primary classification instrument to ensure fair and consistent treatment. This instrument - the COMPAS - is central to the policy's goal of placing inmates in the "least restrictive housing compatible with assessed risks and needs" while ensuring community, staff, and inmate safety. The policy explicitly requires use of a "reliable objective classification tool" for critical decisions such as determining cellmate compatibility for high-risk and special management populations.

Importantly, the policy distinguishes between two classification stages that occur at different points in an inmate's detention. Initial classification occurs at intake/booking and is conducted by booking staff to determine temporary cell assignment, supervision level, and immediate emergency needs; this process does not involve the COMPAS instrument. Primary classification, by contrast, occurs within 72 hours of admission and is conducted by trained classification staff using the COMPAS. Only inmates who remain in custody long enough to complete primary classification receive a COMPAS assessment and are subsequently moved to general housing areas. This procedural distinction has direct implications for

validation research: individuals released quickly, whether through bond, dismissal of charges, or short sentences, never receive a COMPAS score. Consequently, a validation study sample will necessarily be smaller than total booking volume for the same period, reflecting only those detained long enough to undergo primary classification.

The policy framework creates specific validation requirements for the COMPAS tool. Classification staff are permitted to override COMPAS-recommended security designation when aggravating or mitigating circumstances warrant, with 14 approved override reasons documented, ranging from assaultive threats toward staff to developmental disability. These override provisions, combined with the mandate for annual review of the classification plan by leadership and the Classification Committee, establish an accountability structure that necessitates periodic validation of the COMPAS to ensure it accurately predicts institutional risk and appropriately distributes inmates across custody levels.

Methods

Similarities to, and Differences with, the 2020 Validation Study

Similar to the approach used by Jackson and Mendoza (2020) in their initial validation study of COMPAS within the Bernalillo County MDC, we structured our analysis as a retrospective cohort study to explore how well COMPAS risk classification categories (Minimum, Medium, and Maximum) predicted instances of institutional misconduct. Our analytic sample included bookings from April 2021 through December 2024, with specific exclusions described below.

Our revalidation study departs from Jackson and Mendoza's (2020) validation study in several ways. We did not focus our analysis on override use and resultant classification because the observed override rate for the study period (approximately 4.8%) falls within National Institute of Corrections guidelines, overrides are appropriately excluded from predictive validity analyses since they represent instances where classification officer judgment or jail policy superseded the COMPAS recommendation, and we lack sufficiently detailed data on the security-level designation of housing placements. Additionally, evaluating the flow of classification and reclassification patterns does not contribute to an analysis of the COMPAS' predictive power, since COMPAS generates consistent risk scores regardless of classification status, unlike other correctional environments where initial and reclassification tools use different scoring algorithms or weight risk factors differently, conditional on the timing of the assessment administration. We also extended Jackson and Mendoza's (2020) analysis in several ways. We evaluated multiple validation metrics given recent literature highlighting the limitations of relying solely on singular metrics as the primary measure of predictive validity (Moore et al., 2025). To this end, we report survival analyses, AUC results, and odds ratios from logistic regressions to provide a comprehensive perspective on the scope and boundary conditions of the COMPAS' predictive validity. We also explored how well the COMPAS predicts institutional misconduct of any type versus serious misconduct specifically, recognizing that these outcomes may have distinct origins and require separate validation. Finally, we examined the concentration of institutional misconduct, extending research on criminal career concentration that has primarily focused on post-release recidivism rather than within-facility behavior patterns.

Research Design

Following Jackson and Mendoza (2020), we used a booking-level approach, treating each booking episode as a distinct observation. Individuals could receive multiple classifications throughout their detention, either through reclassification while remaining continuously detained or through separate booking episodes. We calculated exposure periods for each booking using documented release dates when available, and we attributed infractions only to the specific booking period during which they occurred, ensuring that each booking's predictive performance was evaluated against the appropriate temporal window.

We applied several exclusion criteria to ensure data quality. We excluded bookings that overlapped with months having incomplete misconduct data (January through March 2021, and July through September 2021) because incomplete recording during these periods could bias estimates of COMPAS predictive validity. We also excluded bookings with release dates after December 31, 2024 since the misconduct data we received only spanned through December 31, 2024. As mentioned, we also excluded overridden cases from the primary analysis because these represented instances where classification officer judgments or mandatory departmental policy superseded COMPAS recommendations, making it inappropriate to evaluate COMPAS's predictive validity for these placements.

We engaged in substantial data cleaning to prepare the dataset for analysis. We standardized data across multiple years and sources to address considerable variation in how dates, names, and other variables were recorded in different files, and we developed parsing algorithms for date variables because some files used standard formats while others required extracting temporal information from filenames or worksheet names. We standardized gender coding because entries appeared in various formats that would have been treated as different categories without harmonization.

We created multiple outcome measures to examine whether COMPAS performed differently across different definitions of institutional misconduct. Our primary outcome was a binary indicator of any infraction during each booking period, capturing whether any misconduct event was recorded regardless of adjudication outcome. We further examined major violations, defined using the MDC's correctional charge codes (e.g., 200-series offense codes including violence, sexual misconduct, weapons, escape-related offenses, and other serious violations), to test whether COMPAS showed differential validity for serious misconduct. For count-based analyses, we used the total number of unique infraction incident days per booking as the outcome. We matched misconduct to specific booking periods by ensuring that incident dates fell between the booking date and the release date.

We used negative binomial regression to model infraction counts, with the natural logarithm of days incarcerated as an offset to account for varying lengths of stay and to model infraction rates per unit time rather than raw counts (Cameron & Trivedi, 2013). We fit models with security level as the primary predictor and extended models including interactions with demographic characteristics (gender and age group), calculating estimated marginal means to obtain adjusted predictions expressed as expected infractions per 1,000 bookings per 30 days (Lenth, 2024). We also fit logistic regression models with a security level by length of stay interaction to estimate predicted probabilities of any infraction across the range of exposure times. We conducted survival analysis using Kaplan-Meier curves and log-rank tests to examine time to first infraction, calculating cumulative incidence estimates at 14, 30, 60, and 90 days to characterize how quickly misconduct occurred within each security classification. We evaluated discriminative ability using receiver operating characteristic (ROC) curve analysis with area under the curve (AUC) calculations, treating security level as an ordinal predictor and testing for differences across demographic subgroups. We examined the concentration of infractions using Pareto analysis, calculating the cumulative proportion of total infractions accounted for by the highest-frequency offenders and computing Gini coefficients to quantify inequality in infraction distribution. Finally, we assessed tool reliability by examining override rates across years and the distribution of time from booking to classification.

Results

Exclusion Criteria

We applied two exclusion criteria to ensure valid observation windows for measuring institutional misconduct. First, we excluded bookings whose detention stay overlapped with months for which

misconduct data were unavailable (January through March 2021 and July through September 2021), as we could not determine whether individuals detained during these periods engaged in misconduct (Maltz, 1999) and thus inclusion of bookings which occurred in this window would lead to an underestimate of misconduct for these specific cases. This criterion excluded 6,784 bookings (i.e., 14.4% of the original sample). Second, we excluded 2,005 bookings lacking valid release dates, as these cases could not be assigned valid exposure time for rate calculations. Missing release dates were concentrated among late 2024 bookings, reflecting individuals who remained in custody at the end of the period for misconduct data collection. After applying both exclusion criteria, our final sample comprised 40,170 booking episodes with complete data on security classification, release dates, and misconduct outcomes.

Characteristics of Sample and Bookings

Our final sample consisted of 18,916 unique individuals incarcerated at the MDC between April 2021 and December 2024. The sample was predominantly male (74.7%, $n = 14,139$), with a mean age at first booking of 39 years ($SD = 11$; median = 37, $IQR = 31-45$ ¹). Approximately half of the sample fell within the 25–39 age range (50.8%, $n = 9,617$), followed by those aged 40 years or older (42.5%, $n = 8,048$) and those aged 17–24 (6.6%, $n = 1,251$). Individuals had an average of 2.12 bookings during the study period ($SD = 2.0$; median = 1, $IQR = 1-2$, range = 1–35). Total days incarcerated had a mean of 70 days ($SD = 137.6$) and a median of 10 days ($IQR = 2-62$).

Table 1. Person-Level Characteristics ($n = 18,916$)

Characteristic	Value
Total Unique Individuals	18,916
Gender, n (%)	
Male	14,139 (74.7%)
Female	4,777 (25.3%)
Age at First Booking, Mean (SD)	39 (11)
Age at First Booking, Median [IQR]	37 [31, 45]
Age Group, n (%)	
17-24 years	1,251 (6.6%)
25-39 years	9,617 (50.8%)
40+ years	8,048 (42.5%)
Bookings per Person, Mean (SD)	2.12 (2)
Bookings per Person, Median [IQR]	1 [1, 2]
Bookings per Person, Range	1 - 35
Total Days Incarcerated, Mean (SD)	70 (137.6)
Total Days Incarcerated, Median [IQR]	10 [2, 62]

Note: IQR = Interquartile Range; SD = Standard Deviation.

We analyzed 40,170 booking episodes across the study period. Males accounted for 75.5% ($n = 30,315$) of bookings, with a mean age at booking of 38.4 years ($SD = 10.2$). Most bookings received medium security designations (54.5%, $n = 21,902$), followed by minimum security (40.7%, $n = 16,356$), with 4.8% ($n = 1,912$) classified as maximum security. Bookings were distributed across study years, with 2021 representing a smaller proportion (12.9%, $n = 5,171$) due to exclusion of bookings overlapping months with incomplete misconduct data, while 2022 through 2024 each contributed approximately 26–31% of the sample. Length of stay showed considerable right skew, with a mean of 33 days ($SD = 78.2$) and a median of 6 days ($IQR = 2-22$); over half of bookings (51.4%, $n = 20,628$) resulted in stays of fewer than

¹ The interquartile range (IQR) represents the middle 50% of observations, spanning from the 25th percentile to the 75th percentile of the distribution.

7 days. Overall, 14.4% (n = 5,768) of bookings involved at least one documented infraction during the detention stay.

Table 2. Booking-Level Characteristics (n = 40,170)

Characteristic	Value
Total Bookings	40,170
Gender, n (%)	
Male	30,315 (75.5%)
Female	9,855 (24.5%)
Age at Booking, Mean (SD)	38.4 (10.2)
Age at Booking, Median [IQR]	37 [31, 44]
Age Group, n (%)	
17-24 years	2,386 (5.9%)
25-39 years	21,623 (53.8%)
40+ years	16,161 (40.2%)
Security Classification, n (%)	
Minimum	16,356 (40.7%)
Medium	21,902 (54.5%)
Maximum	1,912 (4.8%)
Booking Year, n (%)	
2021	5,171 (12.9%)
2022	12,070 (30%)
2023	12,483 (31.1%)
2024	10,446 (26%)
Length of Stay (days), Mean (SD)	33 (78.2)
Length of Stay (days), Median [IQR]	6 [2, 22]
Length of Stay (days), Range	0 - 1069
Length of Stay Category, n (%)	
<7 days	20,628 (51.4%)
7-13 days	6,294 (15.7%)
14-29 days	4,851 (12.1%)
30-89 days	4,440 (11.1%)
90+ days	3,957 (9.9%)
Had Any Infraction, n (%)	5,768 (14.4%)
Had Major Violation, n (%)	N/A
Total Infractions per Booking, Mean (SD)	0.27 (0.94)
Total Infractions per Booking, Median [IQR]	0 [0, 0]

Most individuals had only a single booking (58.1%, n = 10,997), and an additional 17.2% (n = 3,248) had two bookings, meaning approximately three-quarters of the sample had two or fewer bookings during the observation period. The distribution exhibited a right-skewed pattern, with 93.2% of individuals having five or fewer bookings. A small subset of individuals, however, accounted for a disproportionate share of facility admissions: 1.2% (n = 231) had ten or more bookings, with a maximum of 35 bookings for a single individual. This pattern aligns with prior research suggesting that a small group of high-frequency individuals cycles repeatedly through local detention facilities while most have limited contact with the system (Subramanian et al., 2015).

Table 3. Distribution of Bookings per Person

Number of Bookings	N Individuals	Percent	Cumulative %
1	10,997	58.1%	58.1%
2	3,248	17.2%	75.3%
3	1,742	9.2%	84.5%
4	974	5.1%	89.7%
5	662	3.5%	93.2%
6	437	2.3%	95.5%
7	289	1.5%	97.0%
8	207	1.1%	98.1%
9	129	0.7%	98.8%
10+	231	1.2%	100.0%

Institutional Misconduct

Characteristics of Institutional Misconduct

We examined the distribution and characteristics of infractions to understand the nature of institutional misconduct at the MDC. Of the 10,659 total misconduct events recorded during the study period, 8,504 (79.8%) resulted in guilty or sustained findings, while 2,155 (20.2%) were classified as not guilty or dismissed. When we categorized infractions by severity based on guilty/sustained findings only, we found that the majority were classified as major violations (72.6%, $n = 6,177$), followed by institutional infractions (15.9%, $n = 1,355$) and minor infractions (11.4%, $n = 971$). We coded severity based on offense code ranges: 100-series codes (102–120) were classified as institutional or administrative violations (e.g., count violations, failure to follow facility procedures), 200-series codes (201–230) as major violations (e.g., physical violence, sexual misconduct, contraband, drug violations, weapons possession, escape-related offenses), and 300-series codes (301–323) as minor violations (e.g., hygiene infractions, dress code violations, unauthorized exchanges, work performance issues).

Table 4. Infraction Characteristics Overview

Characteristic	Value
Total misconduct events recorded	10,659
Guilty/Sustained Infractions, n (%)	8,504 (47.9%)
Not Guilty/Dismissed, n (%)	9,236 (52.1%)
Infraction Severity, n (%)	
Major	6,177 (72.6%)
Minor	971 (11.4%)
Institutional	1,355 (15.9%)
Unknown	1 (0.0%)

Note: Severity coded based on offense code ranges: 100-series (102-120) = Institutional/Administrative violations, 200-series (201-230) = Major violations, 300-series (301-323) = Minor violations. Severity percentages based on guilty/sustained infractions only.

We ranked infraction types by frequency to identify the most common forms of institutional misconduct among guilty/sustained infractions. Disobedience and insubordination (code 222) was the most prevalent category, comprising 18.8% ($n = 1,597$) of all infractions, followed by disruptive conduct (code 224; 9.1%; $n = 776$), administrative violations (code 111; 6.0%; $n = 511$), verbal abuse or threats (code 221; 5.8%; $n = 496$), and unauthorized areas (code 213; 5.2%; $n = 443$). Violence and physical harm were observed across multiple offense codes (codes 201, 225, and 202), collectively accounting for 11.6% of

all infractions ($n = 987$). The top five offense categories accounted for 45.0% of all infractions, and the top fifteen accounted for over three-quarters (77.2%) of all misconduct incidents. This concentration suggests that a relatively small number of offense types drive most of the recorded misconduct at the MDC.

Table 5. Top 15 Most Common Offense Types

Rank	Code	Description	Count	%	Cumulative %
1	222	Disobedience/Insubordination	1,597	18.8%	18.8%
2	224	Disruptive conduct	776	9.1%	27.9%
3	111	Administrative	511	6.0%	33.9%
4	221	Verbal abuse/Threats	496	5.8%	39.7%
5	213	Unauthorized areas	443	5.2%	45.0%
6	201	Violence/Physical harm	437	5.1%	50.1%
7	210	Program violations	368	4.3%	54.4%
8	312	Dress code	309	3.6%	58.1%
9	215	Contraband	307	3.6%	61.7%
10	225	Violence/Physical harm	305	3.6%	65.3%
11	118	Administrative	279	3.3%	68.5%
12	202	Violence/Physical harm	245	2.9%	71.4%
13	220	Drug violations	201	2.4%	73.8%
14	216	Property violations	159	1.9%	75.6%
15	217	Theft/Stealing	136	1.6%	77.2%

We also calculated infraction rates by gender and age group to explore demographic patterns in institutional misconduct. The overall rate of any misconduct for the full sample was 14.4% ($n = 40,170$; 95% CI: 14.0%–14.7%). Infraction rates were nearly identical between females (14.6%; $n = 9,855$; 95% CI: 13.9%–15.3%) and males (14.3%; $n = 30,315$; 95% CI: 13.9%–14.7%), suggesting minimal gender differences in misconduct prevalence within this population. Age-related patterns were more pronounced: individuals aged 25–39 had the highest infraction rate (16.0%; $n = 21,623$; 95% CI: 15.5%–16.5%), followed by those aged 17–24 (14.5%; $n = 2,385$; 95% CI: 13.1%–16.0%), while individuals older than 39 had the lowest rate (12.1%; $n = 16,161$; 95% CI: 11.6%–12.6%). The finding that older individuals had lower infraction rates is consistent with criminological research indicating that age is inversely associated with rates of institutional misconduct (Steiner & Wooldredge, 2008).

Table 6. Infraction Rates by Demographic Subgroups

Subgroup	N	Any Misconduct	Guilty/Sustained	Not Guilty/Dismissed
Overall	40,170	14.4% (14.0-14.7)	8.9% (8.7-9.2)	8.6% (8.4-8.9)
Gender				
Female	9,855	14.6% (13.9-15.3)	8.4% (7.8-8.9)	9.1% (8.6-9.7)
Male	30,315	14.3% (13.9-14.7)	9.1% (8.8-9.4)	8.5% (8.2-8.8)
Age Group				
17-24	2,385	14.5% (13.1-16.0)	9.4% (8.3-10.6)	9.2% (8.1-10.4)
25-39	21,623	16.0% (15.5-16.5)	9.9% (9.5-10.3)	9.9% (9.5-10.3)
>39	16,161	12.1% (11.6-12.6)	7.6% (7.2-8.0)	6.9% (6.5-7.3)

The Role of Exposure Time

We examined misconduct base rates across categories of length of stay to understand how exposure time related to infraction occurrence. As expected, misconduct rates increased substantially with longer detention periods. Among bookings with same-day release (0 days; $n = 742$), only 0.5% had any

infraction recorded, and this rate increased progressively with longer stays: 1.4% for one-day stays (n = 6,682), 3.4% for two to three-day stays (n = 7,398), 8.1% for four to seven-day stays (n = 7,526), 12.8% for 8–14 days (n = 5,263), 19.8% for 15–30 days (n = 4,370), and 40.0% for stays exceeding 30 days (n = 8,189). The strong positive association between length of stay and misconduct prevalence, with individuals detained more than 30 days having 80 times the misconduct rate of same-day releases, underscores the importance of accounting for exposure time when evaluating misconduct risk, as individuals with shorter stays have fewer opportunities to commit infractions regardless of their underlying risk level.

Table 7. *Misconduct Base Rate by Length of Stay*

Length of Stay	N	N with Infraction	Base Rate (%)
0 days	742	4	0.5
1 day	6,682	91	1.4
2-3 days	7,398	251	3.4
4-7 days	7,526	611	8.1
8-14 days	5,263	672	12.8
15-30 days	4,370	867	19.8
>30 days	8,189	3,272	40.0

Note. Base rate represents the percentage of bookings with at least one infraction within each length of stay category.

We calculated infraction rates per 1,000 person-days per 30 days to standardize comparisons across security classifications with varying lengths of stay. Expressing rates per 1,000 bookings yields whole numbers that are easier to interpret than small decimals, and the 30-day standardization converts daily rates to a monthly timeframe that aligns with typical detention periods. Infraction rates per 1,000 person-days per 30 days generally increased across security classifications over the study period. Among minimum-security bookings, the infraction rate increased from 209.14 in 2021 to 250.68 in 2024; medium-security bookings increased from 215.10 to 312.86 over the same period; and maximum-security bookings increased from 291.48 to 334.86, with a temporary dip to 230.38 in 2022. The percentage of bookings with at least one infraction showed more variable patterns: among minimum-security bookings, this proportion rose from 6.3% in 2021 to 11.3% in 2024, whereas medium-security bookings increased from 13.8% to 17.4%. Interestingly, the percentage of maximum-security bookings with any infraction declined from 24.1% in 2021 to 22.7% in 2024, despite higher infraction rates among those who did engage in misconduct. These trends may reflect changes in facility conditions, enforcement practices, population composition, or other factors that warrant further investigation.

Table 8. *Rate of Misconduct by Year and Security Level*

Year	Security Level	N Bookings	N with Infraction	% with Infraction	Rate per 1,000 per 30 Days
2021	Minimum	2,295	144	6.3%	209.14
2021	Medium	2,627	363	13.8%	215.10
2021	Maximum	249	60	24.1%	291.48
2022	Minimum	4,733	419	8.9%	216.56
2022	Medium	6,791	1,085	16.0%	202.36
2022	Maximum	546	165	30.2%	230.38
2023	Minimum	4,951	559	11.3%	244.17
2023	Medium	6,873	1,223	17.8%	248.76
2023	Maximum	659	179	27.2%	301.17
2024	Minimum	4,377	493	11.3%	250.68
2024	Medium	5,611	974	17.4%	312.86
2024	Maximum	458	104	22.7%	334.86

Note: Rate per 1,000 per 30 Days = (Total infractions / Total person-days) × 1,000 × 30. Analysis restricted to bookings where both admission and release occurred during months with complete misconduct data. Months excluded due to incomplete data: January – March 2021; July – September 2021.

Predictive Validity of COMPAS Security Classifications

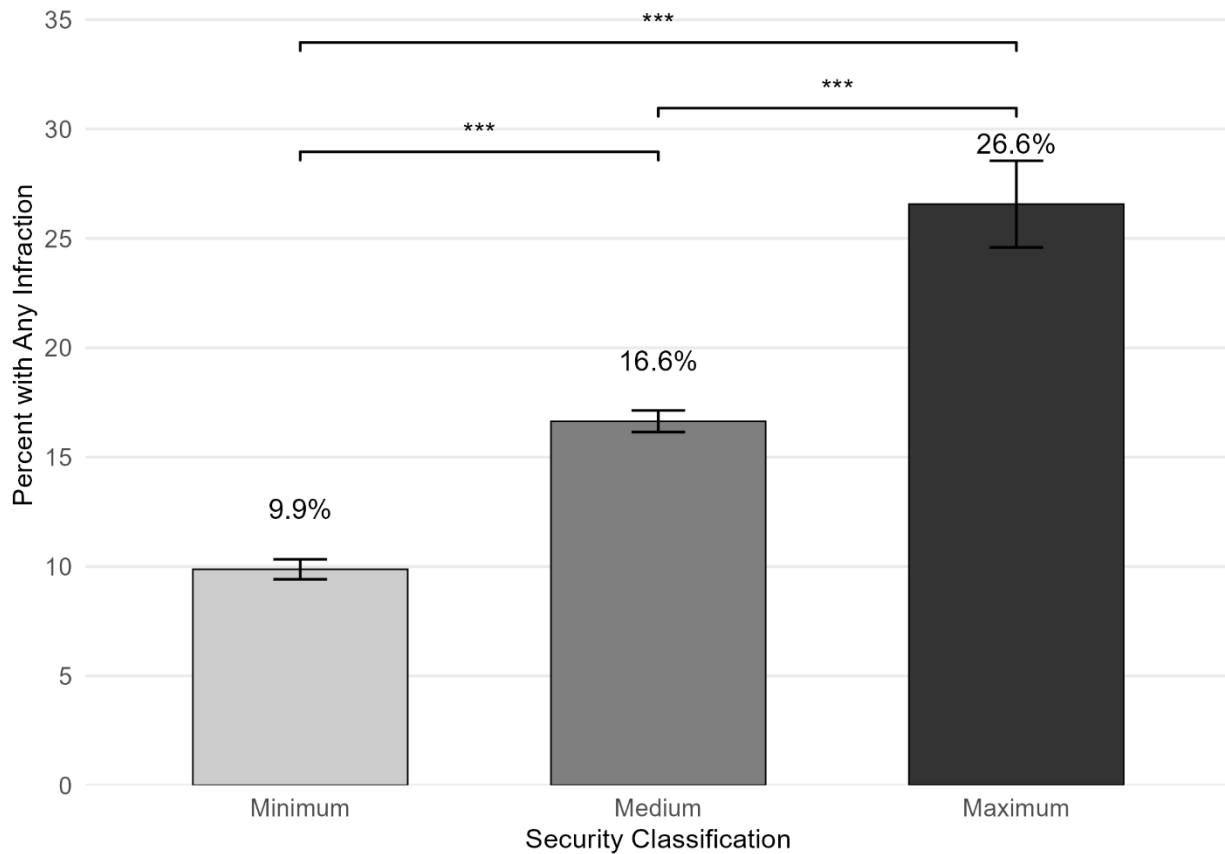
We examined the distribution of infraction counts across COMPAS security classification levels to assess whether the tool captures variation in both the prevalence and intensity of institutional misconduct. The distributions differed significantly by security level, $\chi^2(12) = 643.76$, $p < .001$, with all pairwise comparisons of infraction rates reaching statistical significance after Bonferroni correction ($p < .001$). We found that minimum-security bookings had the highest concentration at zero infractions (90.1%; $n = 14,741$), followed by medium-security (83.4%; $n = 18,257$) and maximum-security bookings (73.4%; $n = 1,404$). When standardized for exposure time, observed misconduct rates showed a clear gradient: minimum-security bookings had 2,802 total infraction events and a rate of 235.2 per 1,000 person-days per 30 days, medium-security bookings had 6,775 events and a rate of 240.4 per 1,000 person-days per 30 days, and maximum-security bookings had 1,111 events and a rate of 275.6 per 1,000 person-days per 30 days. Among bookings with at least one infraction, minimum-security bookings rarely had more than one or two misconduct days, whereas maximum-security bookings exhibited a more dispersed distribution across higher infraction categories: 14.9% had exactly one infraction ($n = 285$), with meaningful proportions extending into higher categories including two infractions (5.0%; $n = 96$), three infractions (2.4%; $n = 46$), four to five infractions (2.5%; $n = 47$), and six or more infractions (1.8%; $n = 34$). In contrast, only 6.7% of minimum-security bookings ($n = 1,091$) had one infraction, with progressively smaller proportions in higher categories.

Table 9. *Distribution of Infraction Counts by Security Classification*

Number of Infractions	Minimum	Medium	Maximum
0	14,741 (90.1%)	18,257 (83.4%)	1,404 (73.4%)
1	1,091 (6.7%)	2,306 (10.5%)	285 (14.9%)
2	279 (1.7%)	657 (3.0%)	96 (5.0%)
3	104 (0.6%)	286 (1.3%)	46 (2.4%)
4-5	85 (0.5%)	259 (1.2%)	47 (2.5%)
6-10	48 (0.3%)	111 (0.5%)	26 (1.4%)
>10	8 (0.0%)	26 (0.1%)	8 (0.4%)
Total	16,356	21,902	1,912

Note: Values are n (%). Infractions represent unique infraction incident days per booking. Chi-square test for independence: $\chi^2(12) = 643.76, p < .001$. All pairwise comparisons of infraction rates were statistically significant ($p < .001$, Bonferroni-corrected).

Figure 1. *Distribution of Infractions by Security Level*



Percentage of bookings with at least one infraction by COMPAS security classification. Error bars represent 95% confidence intervals. *** $p < .001$ (Bonferroni-corrected pairwise proportion tests).

We also estimated marginal predicted values using negative binomial regression with an offset for exposure time, yielding expected infraction rates per 1,000 bookings per 30 days for each security classification level. Individuals classified as minimum security had an expected rate of 235.0 infractions per 1,000 person-days per 30 days (SE = 4.44, 95% CI [226.31, 243.72]). Medium security classifications were associated with a similar expected rate of 240.4 (SE = 2.92, 95% CI [234.65, 246.10]), suggesting that the daily rate of infractions differs only modestly between these two classification levels after accounting for exposure time. Maximum security classifications showed the highest expected rate at 275.6 (SE = 8.27, 95% CI [259.41, 291.82]), representing approximately 1.17 times the rate observed among minimum security bookings. These findings suggest that while the COMPAS classification system distinguishes between risk levels, the exposure-adjusted rate differences between minimum and medium security are relatively modest, with the most pronounced differences appearing between maximum security and the lower classification levels.

Table 10. Observed Infraction Rates by Security Classification Level

Security Level	N	N with Misconduct	Binary Rate (95% CI)	Total Events	Rate per 1,000 per 30 Days	Mean LOS (days)
Minimum	16,356	1,615	9.9% (9.4%-10.3%)	2,802	235.2	21.9
Medium	21,902	3,645	16.6% (16.1%-17.1%)	6,775	240.4	38.6
Maximum	1,912	508	26.6% (24.6%-28.5%)	1,111	275.6	63.2

We also conducted a survival analysis to explore the time to first infraction across security classifications. Among minimum security bookings (n = 16,356), 1,615 individuals engaged in misconduct during follow-up, and the median time to event was not reached, indicating that fewer than half of this group engaged in misconduct during the observation period. Medium security bookings (n = 21,902) had 3,645 events with a median time to event of 159 days, while maximum security bookings (n = 1,912) had 508 events and a median time to event of 93 days. The log-rank test indicated significant differences in time to first infraction across security levels, $\chi^2(2) = 106.02$, $p < .001$. The shorter median survival time for maximum security classifications indicates that individuals in this category tended to engage in misconduct sooner than those in lower security classifications, aligning with the expected pattern that higher risk classifications should be associated with faster onset of misconduct events.

Table 11. Kaplan-Meier Survival Analysis: Time to First Infraction by Security Level

Security Level	N	Events	Median Days to Event
Minimum	16,356	1,615	--
Medium	21,902	3,645	159
Maximum	1,912	508	93

Note. We found significant differences in time to first infraction across security levels (log-rank $\chi^2(2) = 106.02$, $p < .001$), with individuals in Maximum security experiencing infractions more quickly than those in lower security levels. Median days represent the time point at which 50% of individuals in each group experienced their first infraction. "--" indicates median not reached (fewer than 50% of individuals experienced an infraction).

We also examined the cumulative probability of experiencing a first infraction at key time points across security classification levels. The cumulative incidence of misconduct increased over time for all security levels, with higher security classifications showing consistently elevated probabilities at each time point. By day 7, 9.2% (95% CI: 8.6%–9.8%) of minimum security bookings had experienced their first infraction, compared to 10.6% (95% CI: 10.2%–11.1%) for medium security and 13.1% (95% CI: 11.4%–14.7%) for maximum security. These differences widened over longer follow-up periods: by day 90,

32.5% (95% CI: 30.6%–34.4%) of minimum security bookings had experienced at least one infraction, compared to 39.7% (95% CI: 38.2%–41.1%) for medium security and 49.4% (95% CI: 45.1%–53.3%) for maximum security. These findings suggest that individuals classified as higher security not only have higher overall infraction rates and engage in more misconduct but also tend to experience their first infraction earlier in their detention period.

Table 12. *Kaplan-Meier Cumulative Incidence Estimates*

Security Level	Day 7	Day 14	Day 30	Day 60	Day 90
Minimum	9.2% (8.6%-9.8%)	13.8% (12.9%-14.6%)	20.0% (18.8%-21.1%)	27.8% (26.2%-29.4%)	32.5% (30.6%-34.4%)
Medium	10.6% (10.2%-11.1%)	14.6% (14.0%-15.2%)	21.5% (20.7%-22.3%)	31.5% (30.4%-32.7%)	39.7% (38.2%-41.1%)
Maximum	13.1% (11.4%-14.7%)	18.7% (16.6%-20.8%)	26.3% (23.6%-28.9%)	40.2% (36.4%-43.7%)	49.4% (45.1%-53.3%)

Note: Values represent cumulative probability of first misconduct (95% CI) at each time point. Cumulative incidence calculated as 1 minus Kaplan-Meier survival estimate.

We also examined the relationship between COMPAS security classification levels and observed rates of serious institutional misconduct, defined as major severity infractions including violence/physical harm, sexual misconduct, weapons/dangerous items, escape-related incidents, drug violations, and other major offenses. The results indicated a gradient in misconduct rates across security levels. Individuals classified as minimum security exhibited the lowest rate of serious misconduct (5.02%, 95% CI: 4.68%–5.35%, $n = 16,356$), followed by those classified as medium security (8.52%, 95% CI: 8.15%–8.89%, $n = 21,902$), and those classified as maximum security, who showed the highest rate (14.44%, 95% CI: 12.86%–16.01%, $n = 1,912$). The non-overlapping confidence intervals across all three security levels suggest statistically significant differences in misconduct rates between classification groups, supporting the predictive validity of the COMPAS security classification instrument for serious institutional misconduct.

Table 13. *Serious Misconduct Rates by COMPAS Security Level*

Security Level	N	Events	Observed Rate	95% CI
Minimum	16,356	821	5.02%	4.68% - 5.35%
Medium	21,902	1,867	8.52%	8.15% - 8.89%
Maximum	1,912	276	14.44%	12.86% - 16.01%

Note: Serious misconduct defined as Major severity codes including Violence/Physical Harm, Sexual Misconduct, Weapons/Dangerous Items, Escape Related, Drug Violations, and other major offenses.

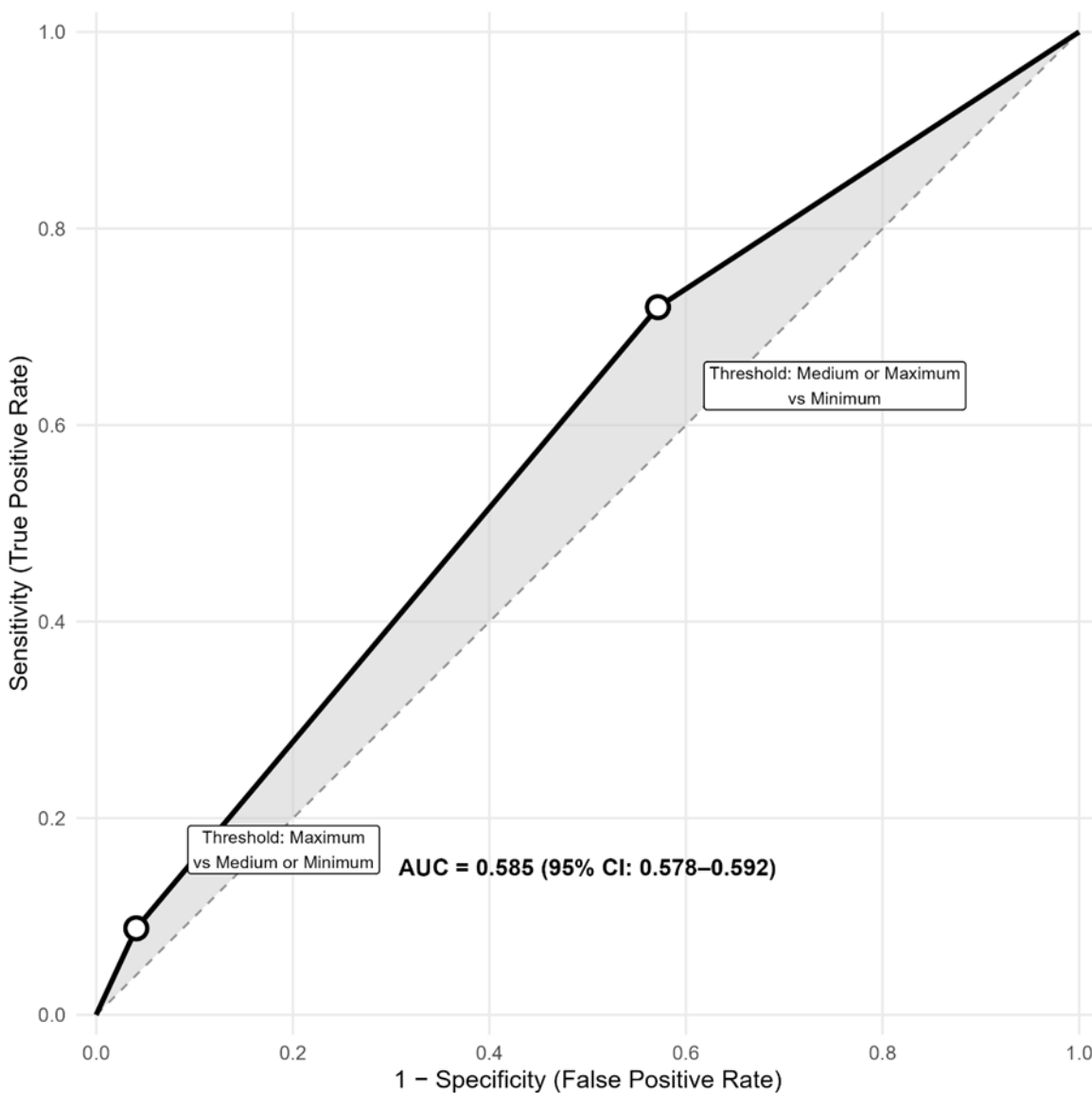
Discrimination Metrics

We assessed COMPAS's discriminative ability² using Receiver Operating Characteristic (ROC) analysis, which evaluates how well a given RAI distinguishes between individuals who did and did not engage in misconduct. The ROC curve plots sensitivity (i.e., the true positive rate) against 1-specificity (i.e., the false positive rate) across all possible classification thresholds, with the area under the curve (AUC) serving as a summary measure of overall discrimination. An AUC of 0.50 indicates performance equivalent to random chance, while an AUC of 1.00 would indicate perfect discrimination. We found that

² Discriminative ability refers to how well a risk assessment tool distinguishes between individuals who do and do not engage in the outcome of interest. A tool with strong discriminative ability will assign higher risk scores to individuals who later engage in misconduct and lower scores to those who do not, effectively separating these two groups.

the COMPAS achieved an AUC of 0.585 (95% CI: 0.576–0.594), suggesting that when we randomly select one individual who engaged in misconduct and one who did not, the COMPAS tool correctly assigns a higher security classification to the individual who engaged in misconduct approximately 59% of the time. It is important to note that this AUC estimate is derived from a trichotomized ordinal predictor with only three categories (Minimum, Medium, Maximum), which likely constrains the maximum achievable AUC and likely underestimates the true discriminative ability of the underlying continuous COMPAS score.

Figure 2. ROC Plot of the COMPAS AUC



ROC curve for COMPAS security classification with annotated classification thresholds.
Shaded area represents the AUC.
Operating points show the sensitivity-specificity trade-off at each possible decision threshold.

According to standards established by Desmarais and Singh (2013) for evaluating the predictive validity of criminological RAIs using AUCs, our observed AUC of 0.585 falls within the "fair" range (0.55–0.63), which is consistent with findings from other validation studies of risk assessment instruments in correctional settings that typically report AUC values between 0.55 and 0.71 for predicting institutional

misconduct (Fazel et al., 2022). However, what qualifies as a "good" AUC score depends on the normative benchmarks used. Equivant (2019) characterizes AUC values between 0.55 and 0.63 as "fair" and values above 0.64 as "good," while Dieterich et al. (2016) suggest that AUC values above 0.70 indicate "good" discrimination. Rice and Harris (2005) proposed an alternative framework based on effect size equivalencies, characterizing AUC values of 0.56 as small, 0.64 as medium, and 0.71 as large effects, placing our observed AUC within the small-to-medium range. Given this variation in how researchers define "goodness" in AUC interpretation, our observed AUC of 0.585 would be characterized as "fair" by both Desmarais and Singh (2013) and Equivant (2019) standards, as a "small" effect size per Rice and Harris (2005), and would fall below the "good" threshold established by Dieterich et al. (2016).

We found that the COMPAS classification system exhibited the sensitivity-specificity trade-off common to RAIs. Using only Maximum classification as the high-risk threshold yielded high specificity (95.9%) but poor sensitivity (8.8%), identifying few individuals who would subsequently have infractions. A broader threshold (Maximum or Medium vs. Minimum) improved sensitivity to 72.0% but reduced specificity to 42.8%. Positive predictive values were modest at both thresholds (17.4%-26.6%), reflecting the relatively low base rate of institutional misconduct. However, negative predictive values were high (86.3%-90.1%), suggesting the COMPAS may be more useful for identifying individuals at low risk for misconduct than for identifying those at high risk.

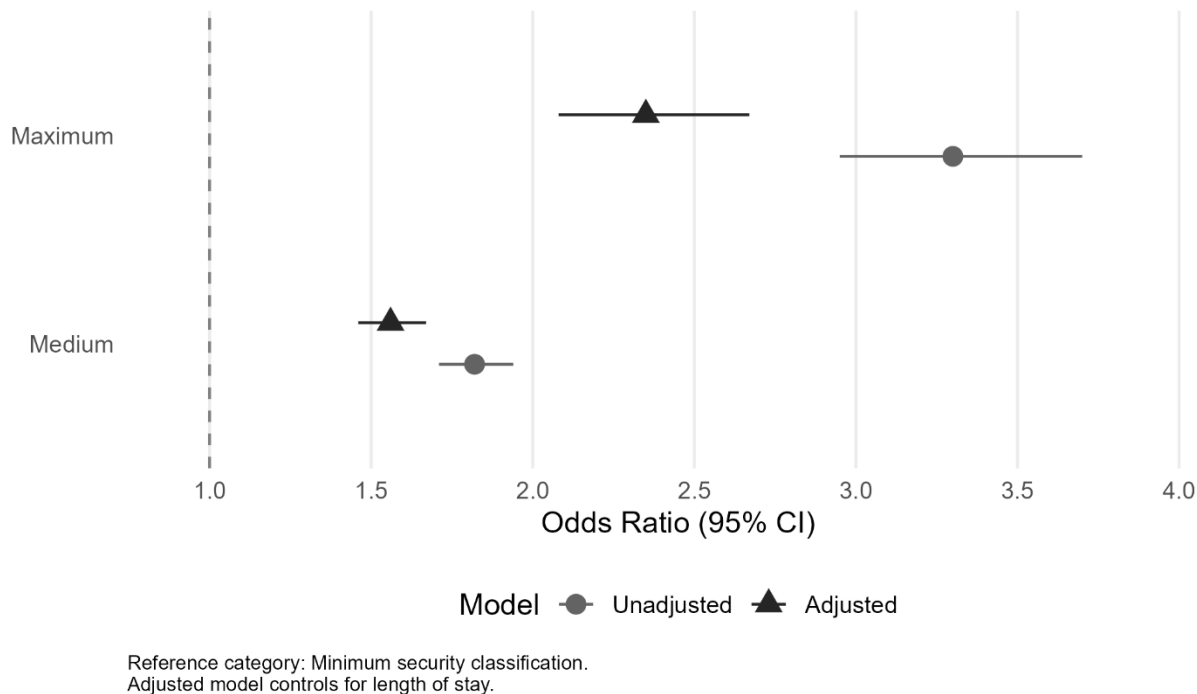
Table 14. *Alternative Discrimination Metrics*

Classification Threshold	Sensitivity	Specificity	PPV	NPV	Accuracy
Maximum vs. Others	8.8%	95.9%	26.6%	86.3%	83.4%
Maximum + Medium vs. Minimum	72.0%	42.8%	17.4%	90.1%	47.0%

Note. PPV = Positive Predictive Value; NPV = Negative Predictive Value. 'Maximum vs. Others' treats only Maximum security as high-risk. 'Maximum + Medium vs. Minimum' treats both Maximum and Medium as high-risk. Trade-off between sensitivity and specificity evident across thresholds.

Importantly, the odds ratios derived from logistic regression models for comparing maximum to minimum security classifications indicate that individuals in maximum security had nearly three times the likelihood of misconduct, demonstrating that the COMPAS appropriately ranks risk, even if the AUC metric does not fully capture this separation due to its compressed scale points. In unadjusted logistic regression models, we found that medium security was associated with 1.82 times the odds of infraction relative to minimum (95% CI: 1.71–1.94), while maximum security was associated with 3.30 times the odds (95% CI: 2.95–3.70). After adjusting for length of stay, these associations attenuated but remained statistically significant, with odds ratios of 1.56 (95% CI: 1.46–1.67) for medium and 2.35 (95% CI: 2.08–2.67) for maximum security relative to minimum.

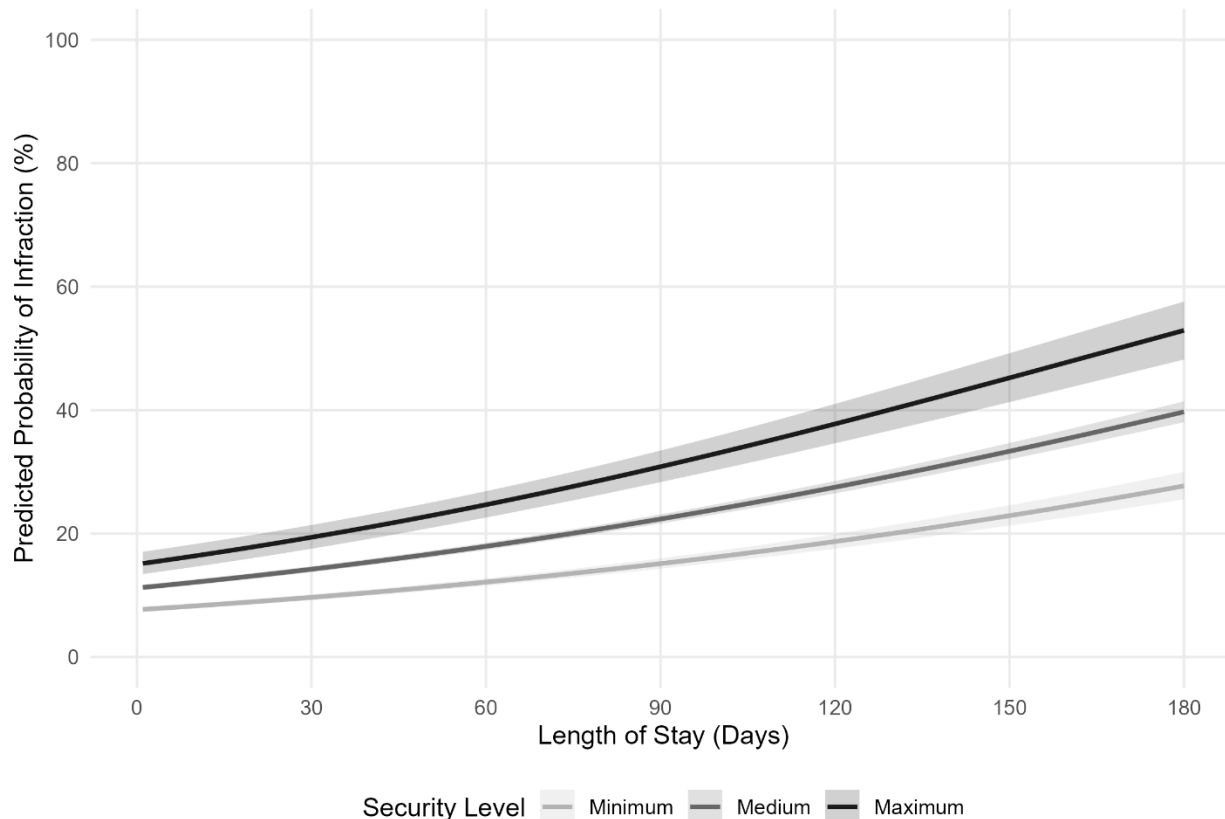
Figure 3. Odds Ratio for Misconduct (Unadjusted and Adjusted for Exposure Time)



We also used logistic regression to model the predicted probability of engaging in misconduct as a continuous function of length of stay, stratified by security classification level. The resulting probability curves illustrate how the risk of infraction accumulates over time and how this accumulation varies across security levels. At 30 days, predicted probabilities were 9.7% (95% CI: 9.2%-10.1%) for minimum security, 14.2% (95% CI: 13.7%-14.7%) for medium security, and 19.4% (95% CI: 17.6%-21.4%) for maximum security. As the length of stay increased, all three security levels showed substantial increases in predicted probability of infraction. By 90 days, predicted probabilities reached 15.1% (95% CI: 14.3%-16.0%) for minimum security, 22.4% (95% CI: 21.6%-23.1%) for medium security, and 30.8% (95% CI: 28.3%-33.4%) for maximum security. At 180 days, predicted probabilities continued to diverge across security levels: 27.7% (95% CI: 25.6%-30.0%) for minimum, 39.7% (95% CI: 38.1%-41.4%) for medium, and 52.9% (95% CI: 48.2%-57.6%) for maximum security.

Unlike patterns we reported on in our prior analyses, we found that the gap between security levels widened rather than converged over time, with maximum-security individuals exhibiting approximately twice the predicted infraction probability of minimum-security individuals at 180 days. These model-based predictions complement our categorical analyses by showing that risk accumulates continuously rather than in discrete steps, and they reinforce the finding that both security classification and length of stay independently contribute to the probability of infraction.

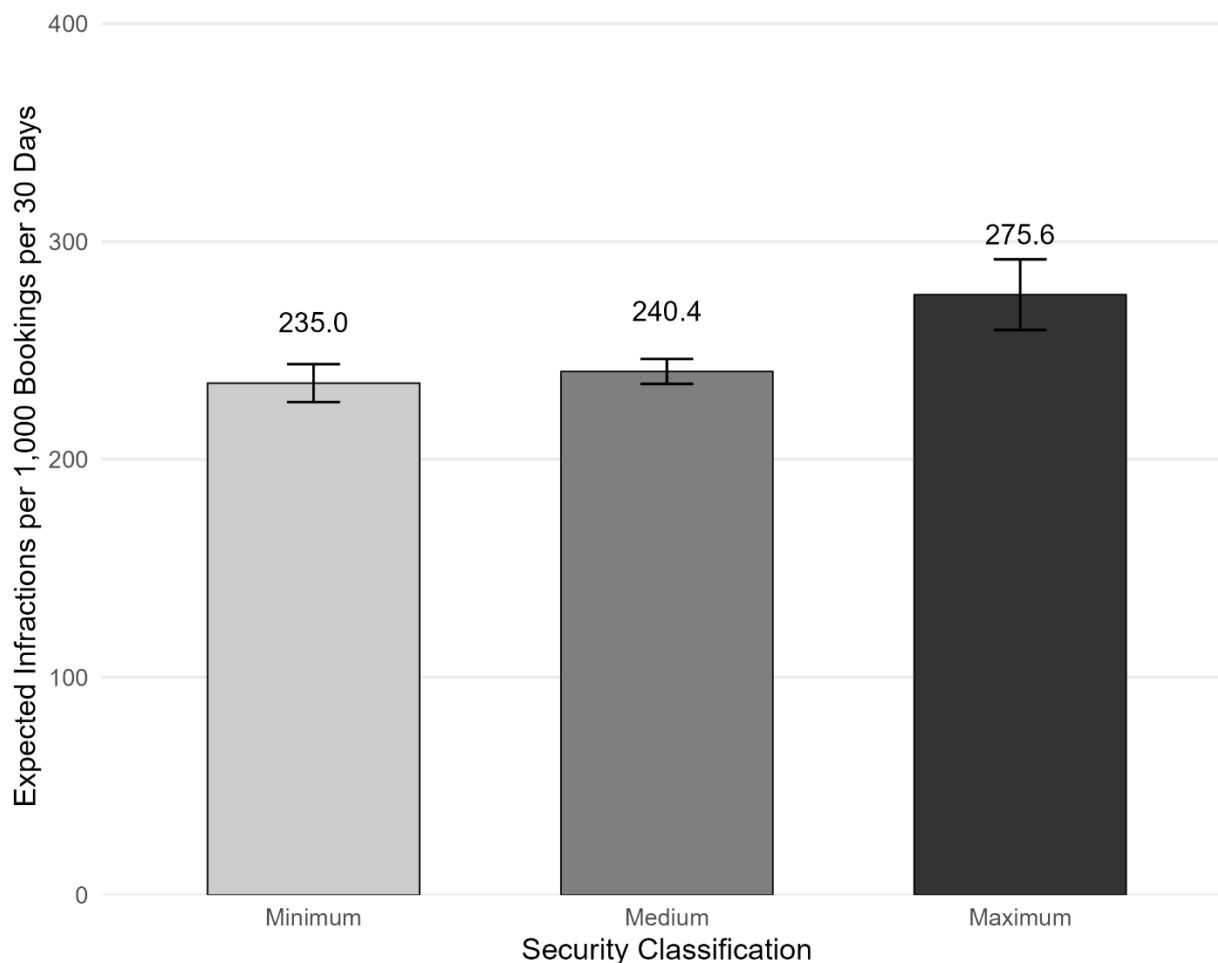
Figure 4. *Predicted Probability of Infraction by Security Level and Length of Stay*



Predicted probability of any infraction by length of stay and COMPAS security classification.
 Shaded regions represent 95% confidence intervals.
 Estimates derived from logistic regression with security level × length of stay interaction.

Following Jackson & Mendoza (2020), we estimated marginal predicted values using negative binomial regression with an exposure-time offset, yielding expected infraction rates per 1,000 bookings over 30 days for each security classification level. Individuals classified as minimum security had an expected rate of 235.0 infractions per 1,000 person-days per 30 days (SE = 4.44, 95% CI [226.31, 243.72]). Medium-security classifications were associated with a similar expected rate of 240.4 (SE = 2.92, 95% CI [234.65, 246.10]), suggesting that the daily rate of infractions differs only modestly between these two classification levels after adjusting for exposure time. The maximum-security classification showed the highest expected rate at 275.6 (SE = 8.27, 95% CI [259.41, 291.82]), approximately 1.17 times the rate observed for minimum-security bookings. These findings suggest that, although the COMPAS classification system distinguishes among risk levels, exposure-adjusted rate differences between minimum and medium security are relatively modest, with the largest differences observed between maximum security and the lower classification levels. Together, these findings provide some support for the discriminant validity of the COMPAS security classification system, as higher-risk classifications were associated with higher rates of institutional misconduct.

Figure 5. *Estimated Infraction Rates per 1,000 Bookings per 30 Days by Security Level*



Estimated infraction rates by COMPAS security classification level.
 Error bars represent 95% confidence intervals.
 Estimates derived from negative binomial regression model with offset for exposure time.

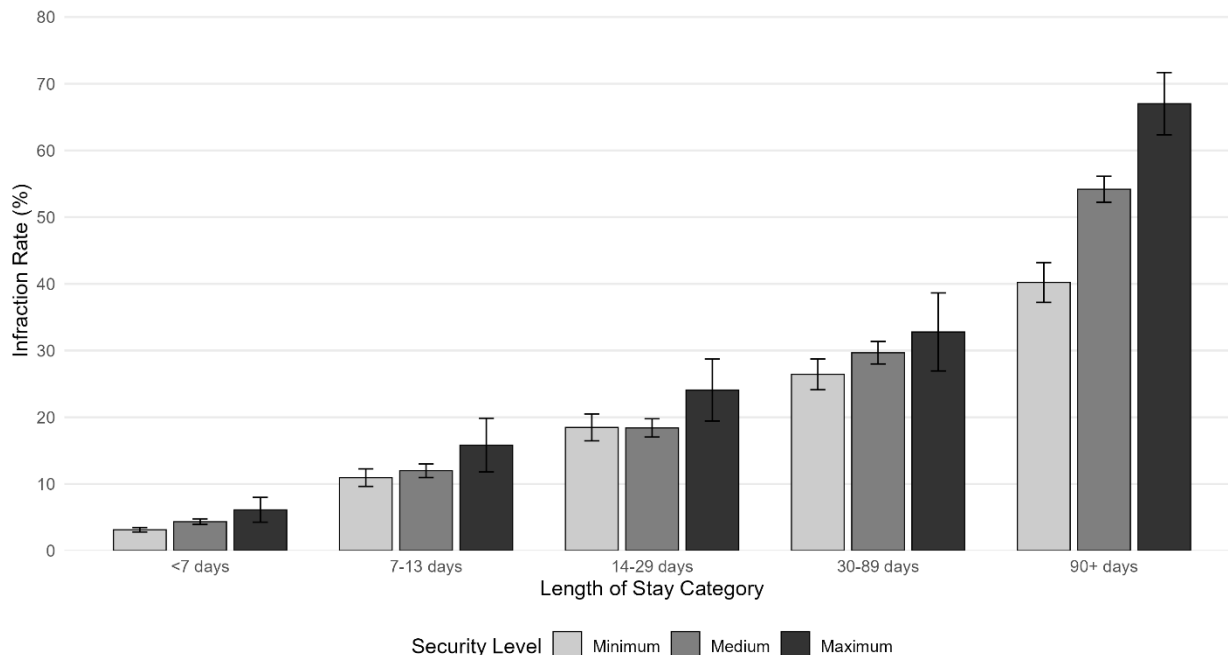
Subgroup Analyses

We also explored how infraction rates varied across length-of-stay categories, stratified by COMPAS security classification level. As shown in Figure 6, infraction risk increased monotonically with longer stays across all three security levels, but the rate of increase differed substantially by classification. For the shortest stays (fewer than 7 days), infraction rates ranged from 3.1% (95% CI: 2.8%-3.5%) for minimum security to 4.3% (95% CI: 3.9%-4.8%) for medium security to 6.1% (95% CI: 4.3%-8.0%) for maximum security. By 90 or more days, individuals classified as maximum security had an infraction rate of 67.0% (95% CI: 62.3%-71.7%), compared to 54.2% (95% CI: 52.2%-56.1%) for medium security and 40.2% (95% CI: 37.2%-43.2%) for minimum security.

Maximum security classifications showed the steepest progression, with infraction rates climbing from approximately 6% at the shortest stays to 67% at the longest stays - more than a tenfold increase. In contrast, minimum security classifications showed a more gradual trajectory, increasing from

approximately 3% to 40% over the same length-of-stay range. The widening gap between security levels at longer stays suggests that COMPAS classifications become increasingly predictive as exposure time increases, which may reflect both higher baseline risk among those classified as maximum security and potentially greater accumulation of risk over time for this group. This pattern is also partly expected given that maximum security individuals have longer average lengths of stay, and because most individuals are released within seven days, those with extended stays represent a more selected subset of the population regardless of security classification.

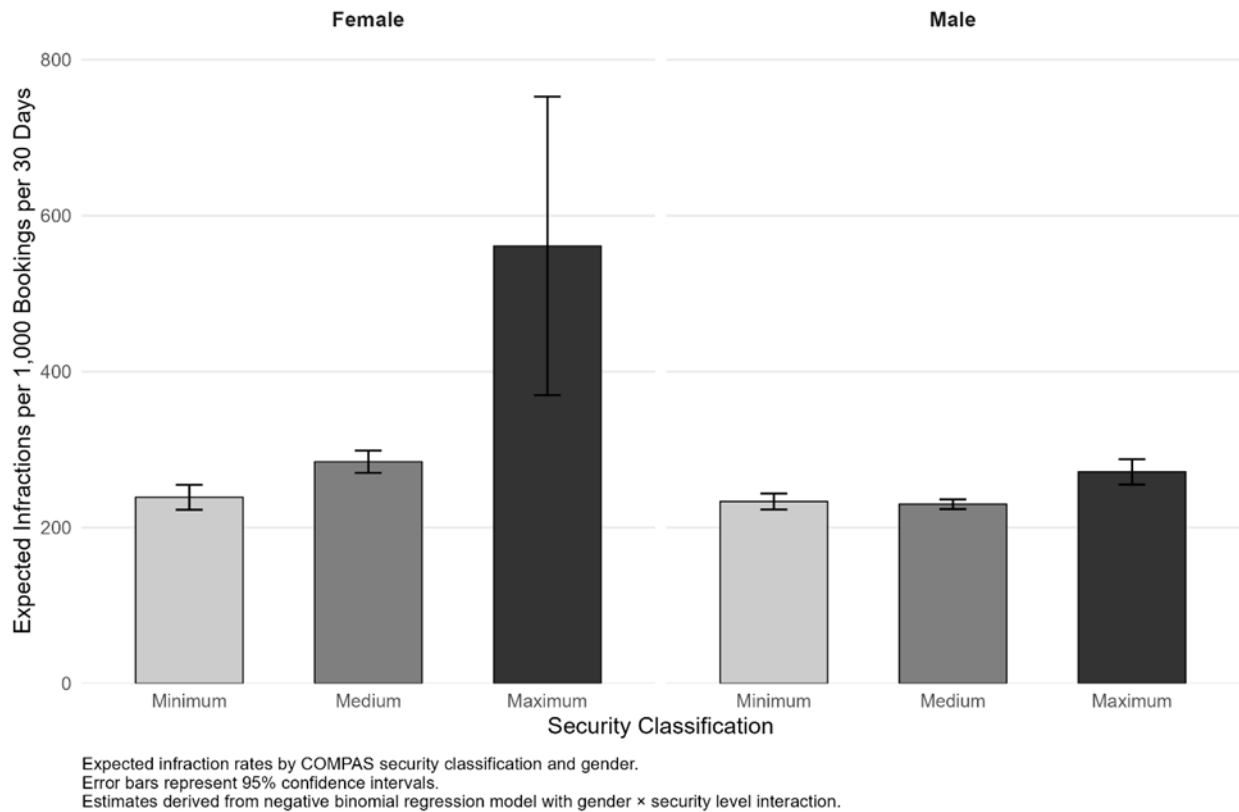
Figure 6. *Infraction Rate by Length of Stay Category x COMPAS Score*



Percentage of bookings with at least one infraction by length of stay category and COMPAS security classification. Error bars represent 95% confidence intervals.

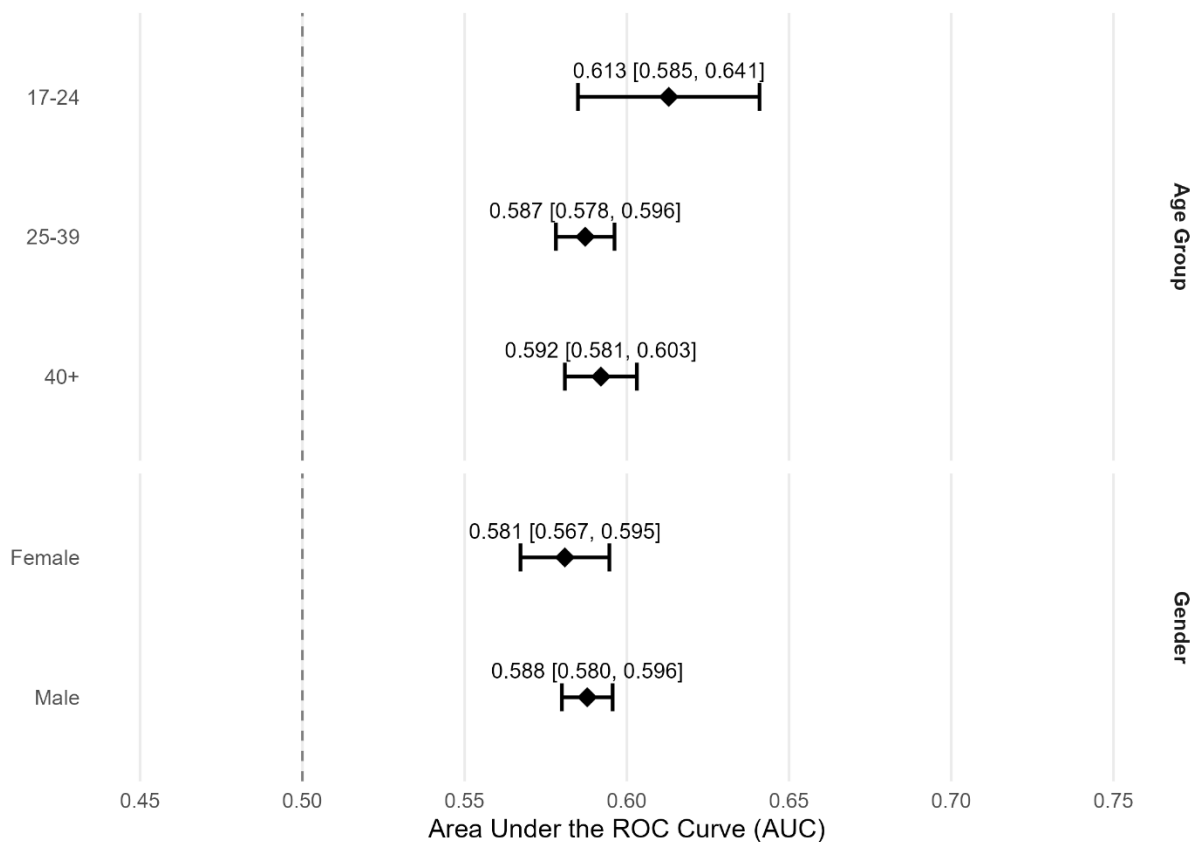
We also stratified expected infraction rates by gender and security classification to explore whether COMPAS classifications perform consistently across male and female populations. At the minimum security level, expected rates were similar for females (238.80 per 1,000 per 30 days, 95% CI [222.77, 254.82]) and males (233.40, 95% CI [223.03, 243.77]), with largely overlapping confidence intervals. However, at the medium security level, females showed notably higher expected rates (284.45, 95% CI [270.21, 298.69]) compared to males (229.96, 95% CI [223.73, 236.18]), with non-overlapping confidence intervals suggesting a statistically meaningful difference. This pattern was most pronounced at the maximum security level, where females had expected rates of 561.22 (95% CI [369.74, 752.71]) compared to 271.39 for males (95% CI [255.19, 287.59]). The substantially wider confidence interval for maximum security females reflects the smaller sample size in this subgroup. These findings suggest that the relationship between security classification and infraction rates may vary by gender, with females classified at higher security levels potentially showing elevated misconduct rates relative to their male counterparts, though it is hard to generalize given the small sample size of female inmates with maximum security level ($n = 22$).

Figure 7. *Expected Infractions per 1,000 Bookings per 30 Days*



Similarly, we stratified expected infraction rates by age group and security classification to explore whether COMPAS classifications perform consistently across different age categories. Similar to our gender analysis, this approach helps determine whether the tool's predictive accuracy differs between younger and older individuals. Younger individuals (aged 17–24) consistently showed the highest expected infraction rates across all security levels. At minimum security, those aged 17–24 had an expected rate of 442.41 (95% CI [364.75, 536.61]), compared to 333.24 for those aged 25–39 (95% CI [310.59, 357.53]) and 150.03 for those over 39 (95% CI [134.08, 167.89]). This age gradient was present across all security classifications, with individuals over 39 showing expected rates generally one-third to one-half those of the youngest age group. Within the maximum security classification, expected rates ranged from 621.87 for the youngest group (95% CI [281.62, 1373.19]) to 327.19 for the oldest group (95% CI [209.57, 510.81]), although confidence intervals were notably wider for maximum-security estimates due to smaller sample sizes.

We also explored whether the discriminative validity of COMPAS security classifications varied across demographic subgroups. AUC values were generally consistent across age groups and gender, with all subgroup estimates falling within the "fair" range and confidence intervals largely overlapping. The youngest age group (17–24) showed the highest AUC at 0.613 (95% CI: 0.585–0.641), followed by those over 40 (AUC = 0.592, 95% CI: 0.581–0.603) and those aged 25–39 (AUC = 0.587, 95% CI: 0.578–0.596). AUC values were nearly identical across gender, with males showing an AUC of 0.588 (95% CI: 0.580–0.596) and females showing an AUC of 0.581 (95% CI: 0.567–0.595). The overlapping confidence intervals across all subgroups suggest that COMPAS performs comparably across demographic categories, with no evidence of substantially diminished predictive validity for any particular age or gender group.

Figure 8. *AUCs by Age and Gender*

AUC values with 95% confidence intervals for COMPAS security classification by demographic subgroup. Dashed vertical line represents chance performance (AUC = 0.50). Confidence intervals computed using DeLong's method.

Concentration of Misconduct

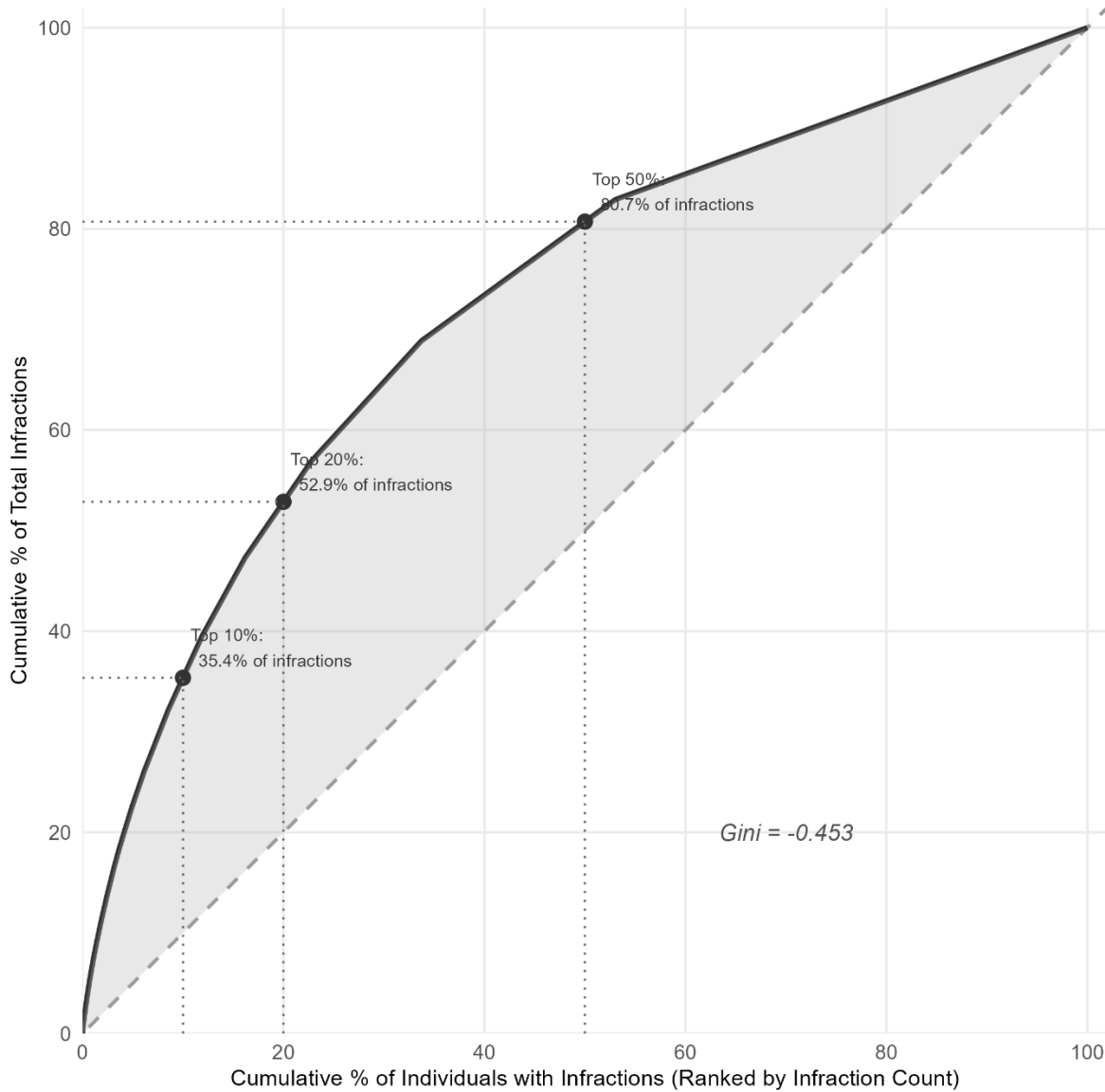
We also examined the distribution of misconduct among individuals with the highest counts to assess whether a small subset of the population accounted for a disproportionate share of misconduct. This helps us evaluate whether targeted interventions for high-risk individuals could meaningfully reduce overall misconduct rates. Among individuals with any infraction, the top 10% ($n = 389$) accounted for 35.4% ($n = 3,780$) of all infractions. When we expanded to the top 20% ($n = 778$ individuals), this group accounted for over half (52.9%, $n = 5,650$) of all infractions. The top 50% of individuals with misconduct ($n = 1,945$) accounted for 80.7% ($n = 8,627$) of all misconduct incidents. This pattern indicates that infractions were highly concentrated among a relatively small group of individuals (Gini coefficient = 0.453), a finding consistent with research on the broader concentration of antisocial behavior (Vaughn et al., 2014). From a practical standpoint, these results suggest that interventions targeting the highest-risk individuals could address a substantial proportion of institutional misconduct.

Table 15. Pareto Analysis - Concentration of Infractions Among High-Frequency Individuals

Percentile	N People	Total Infractions	% of Total Infractions	% of People with Infractions
Top 10%	389	3,780	35.4	10
Top 20%	778	5,650	52.9	20
Top 50%	1,945	8,627	80.7	50

Note: We found a substantial concentration of infractions among a small proportion of individuals, indicating that targeted interventions may have a disproportionate impact on overall facility infraction rates. This pattern suggests that a relatively small number of high-risk individuals account for a large share of institutional misconduct. Analysis restricted to individuals with at least one infraction (n = 3,889). Gini coefficient = -0.453.

Figure 9. Pareto Analysis - Concentration of Infractions Among High-Frequency Individuals



Lorenz curve showing concentration of infractions among individuals with at least one infraction. Dashed diagonal represents perfect equality; shaded area indicates inequality. Higher Gini coefficients indicate greater concentration of infractions among fewer individuals.

Reliability of Tool Implementation

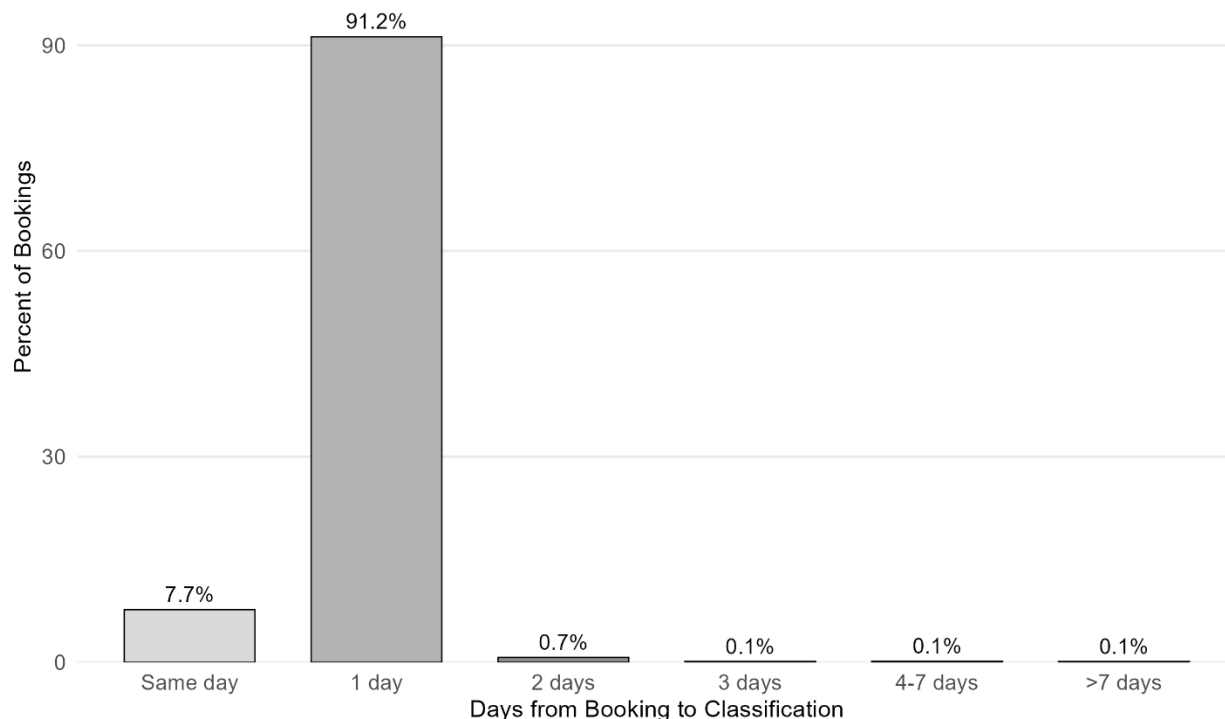
We also examined override rates to assess the consistency with which the COMPAS was applied over time. During the study period (2021–2024), 95.2% (n = 70,465) of primary classifications were not overridden, whereas 4.8% (n = 3,527) were overridden by facility staff. Override rates remained stable throughout the years, ranging from 3.5% to 5.9%, suggesting that staff generally accepted the COMPAS classification recommendations.

Table 16. *Classification Override Rates by Year*

Year	Total Classification Events	No Override	Override
2021	17,995	16,934 (94.1%)	1,061 (5.9%)
2022	20,142	19,246 (95.6%)	896 (4.4%)
2023	18,299	17,351 (94.8%)	948 (5.2%)
2024	17,556	16,934 (96.5%)	622 (3.5%)
Overall	73,992	70,465 (95.2%)	3,527 (4.8%)

We also found that COMPAS assessments were administered promptly after booking: the median time from booking to classification was 1 day (IQR: 1–1), with a mean of 0.97 days (SD = 2.21). Most classifications (97.7%) occurred on the same day as booking, and 98.9% were completed within one day. This timely administration indicates that classification information was available early in the booking period to inform housing and supervision decisions.

Figure 10. *Typical Time to Primary Classification*



Distribution of time between booking and COMPAS classification.
The majority of classifications occurred within one day of booking, indicating timely tool administration.

Discussion

We conducted a validation study of the COMPAS RAI at the Bernalillo County Metropolitan Detention Center using data from April 2021 through December 2024, encompassing 40,170 booking episodes among 18,916 unique individuals. We aimed to provide an independent evaluation of COMPAS's predictive validity for institutional misconduct and to assess the objectivity and reliability of its implementation within the MDC. The findings reported here are intended to inform the federal monitor's ongoing oversight of classification practices at the MDC and to support evidence-based decision-making regarding the continued use of COMPAS for security classification.

Our findings provide support for the predictive validity of COMPAS for institutional misconduct at the MDC. We observed a clear and statistically significant gradient in misconduct rates across security classification levels, with 9.9% of minimum security bookings, 16.6% of medium security bookings, and 26.6% of maximum security bookings having at least one documented misconduct event. This gradient was also evident for serious misconduct specifically, with rates of 5.0%, 8.5%, and 14.4% for minimum, medium, and maximum security, respectively. The pattern held across multiple analytic approaches, including exposure-adjusted negative binomial regression and survival analysis, with survival models showing that maximum security bookings reached median time to first infraction at 93 days compared to 159 days for medium security, whereas minimum security bookings did not reach median (indicating fewer than half engaged in any misconduct). The relationship between COMPAS classifications and misconduct also strengthened as length of stay increased: among bookings with stays under 7 days,

infraction rates ranged from 3.1% to 6.1% across security levels, but by 90 or more days, rates diverged substantially to 40.2%, 54.2%, and 67.0% for minimum, medium, and maximum security, respectively.

The overall AUC for COMPAS security classifications was 0.585 (95% CI: 0.576–0.594), which falls within the "fair" range (0.55–0.63) according to standards established by Desmarais and Singh (2013) and is consistent with AUC values typically reported for risk assessment instruments predicting institutional misconduct (Fazel et al., 2022). Discriminative validity was consistent across demographic subgroups, with AUCs ranging from 0.587 to 0.613 across age groups and from 0.581 to 0.588 across gender, and overlapping confidence intervals suggest that COMPAS performs comparably across these categories. Although the AUC reflects modest discrimination by conventional standards, the exposure-adjusted odds ratios comparing maximum to minimum security classifications indicate that the COMPAS appropriately ranks risk, with maximum-security bookings showing nearly three times the misconduct rate of minimum-security bookings.

A key consideration for the federal monitor is whether COMPAS is being implemented objectively and reliably, and our findings provide evidence supporting both. Override rates remained low and stable across the study period (3.5%–5.9% annually; 4.8% overall), indicating that classification officers generally accepted COMPAS recommendations rather than substituting subjective judgments. The override rate aligns with standards set forth by the National Institute of Corrections. Moreover, COMPAS assessments were administered promptly and consistently, with 98.9% of primary classifications completed within one day of booking, ensuring that classification information was available early to inform housing decisions. The use of a standardized actuarial RAI promotes objectivity by reducing reliance on individual officer judgment, which research has shown can be subject to inconsistency and bias (Viljoen et al., 2025), and the low override rate indicates that the MDC has maintained fidelity to consistently using an RAI.

Several limitations warrant consideration. We relied on documented disciplinary infractions as our outcome measure, which may underestimate the true rate of rule violations due to underdetection, discretionary reporting by correctional staff, and selective enforcement practices that vary across housing units and shifts. Six months of misconduct data were unavailable (January–March 2021 and July–September 2021), necessitating the exclusion of bookings that overlapped with these periods and potentially introducing selection bias if individuals booked during these windows differed systematically from the analytic sample. Our analysis was restricted to a single county detention facility and may not generalize to other correctional settings with different populations, operational practices, or misconduct recording procedures. We also could not establish whether COMPAS classifications have causal effects on subsequent infractions, as individuals classified at higher security levels may receive differential treatment, housing assignments, or programming access that independently influences misconduct risk.

In conclusion, our findings support the continued use of COMPAS as one component of a comprehensive classification approach at the MDC. The consistent gradient in misconduct rates across security levels, even after adjusting for exposure time, indicates that COMPAS classifications provide useful information for risk differentiation, and the low override rates and prompt administration indicate that the tool is being implemented objectively and reliably. At the same time, specific findings warrant ongoing attention: the low sensitivity when using maximum security as the high-risk threshold indicates that supplementary approaches, including behavioral observation during detention, reassessment protocols, and clinical judgment, remain important complements to actuarial classification. We recommend ongoing monitoring of classification performance, including regular reporting to the federal monitor on the metrics examined in our study, to ensure the COMPAS continues to perform as expected and that any changes in predictive validity are detected promptly.

References

- Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Anderson Publishing.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Applegarth, D. M., Lewis, R. A., & Rief, R. M. (2023). Imperfect tools: A research note on developing, applying, and increasing understanding of criminal justice risk assessments. *Criminal Justice Policy Review*, 34(4), 431–446. <https://doi.org/10.1177/08874034231180505>
- Austin, J., Bhati, A., Dedel, K., Rantala, R., & Walsh, A. (2013). *Assessing the performance of risk assessment instruments*. National Institute of Justice.
- Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1), 21–40. <https://doi.org/10.1177/0093854808326545>
- Brookings Institution. (2023, June 27). *Understanding risk assessment instruments in criminal justice*. <https://www.brookings.edu/articles/understanding-risk-assessment-instruments-in-criminal-justice/>
- Bureau of Justice Assistance. (2020). *Risk validation*. <https://bja.ojp.gov/program/psrac/validation/risk-validation>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Desmarais, S. L., & Singh, J. P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Council of State Governments Justice Center.
- Dhami, M. K., & Belton, I. K. (2017). On getting inside the judge's mind. *Translational Issues in Psychological Science*, 3(2), 214–226. <https://doi.org/10.1037/tps0000115>
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), Article eaao5580. <https://doi.org/10.1126/sciadv.aao5580>
- Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *BMJ*, 345, Article e4692. <https://doi.org/10.1136/bmj.e4692>
- Fazel, S., Wolf, A., Larsson, H., Lichtenstein, P., Mallett, S., & Fanshawe, T. R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *BMJ*, 378, Article e070939. <https://doi.org/10.1136/bmj-2022-070939>
- 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.
- Gendreau, P., Goggin, C., & Law, M. (1997). Predicting prison misconduct. *Criminal Justice and Behavior*, 24(4), 414–431. <https://doi.org/10.1177/0093854897024004002>

- Goffman, A. (2009). On the run: Wanted men in a Philadelphia ghetto. *American Sociological Review*, 74(3), 339–357. <https://doi.org/10.1177/000312240907400301>
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293–323. <https://doi.org/10.1037/1076-8971.2.2.293>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Northpointe Inc. (2012). *COMPAS risk & need assessment system: Selected questions posed by inquiring agencies*. Author.
- Schlager, M. D., & Robbins, K. (2008). Does parole work? Revisited: Reframing the discussion of the impact of postprison supervision on offender outcome. *The Prison Journal*, 88(2), 234–251. <https://doi.org/10.1177/0032885508319164>
- Skeem, J., & Lowenkamp, C. (2025). Algorithmic bias in criminal risk assessment: The consequences of racial differences in arrest as a measure of crime. *Annual Review of Criminology*, 8, 97–119. <https://doi.org/10.1146/annurev-criminol-022422-125019>
- Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4), 680–712. <https://doi.org/10.1111/1745-9125.12123>
- Steiner, B., Butler, H. D., & Ellison, J. M. (2014). Causes and correlates of prison inmate misconduct: A systematic review of the evidence. *Journal of Criminal Justice*, 42(6), 462–470. <https://doi.org/10.1016/j.jcrimjus.2014.08.001>
- Steiner, B., & Wooldredge, J. (2015). Individual and environmental sources of work and family stress among prison officers. *Criminal Justice and Behavior*, 42(8), 800–818. <https://doi.org/10.1177/0093854814564463>
- Viljoen, J. L., Jonnson, M. R., Cochrane, D. M., Vargen, L. M., & Vincent, G. M. (2025). Are risk assessment tools more accurate than unstructured judgments in predicting violent, any, and sexual offending? A meta-analysis of direct comparison studies. *Behavioral Sciences & the Law*, 43(1), 3–40. <https://doi.org/10.1002/bsl.2698>