Document Title:	**AI R&D to Support Community Supervision: Integrated Dynamic Risk Assessment for Community Supervision (IDRACS), Final Report**
Author(s):	**Pamela K. Lattimore, Ph.D., Christopher Inkpen, Ph.D.**
Document Number:	**309339**
Date Received:	**July 2024**
Award Number:	**2019-75-CX-0012**

AI R&D to Support Community Supervision: Integrated Dynamic Risk Assessment for Community Supervision (IDRACS)

Final Report

Pamela K. Lattimore, Ph.D.
Principal Investigator
lattimore@rti.org
1.919.485.7759

Christopher Inkpen, Ph.D.
Project Director
cinkpen@rti.org
1.919.541.6411

RTI International
3040 E. Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709
www.rti.org

NIJ Grant Number 2019-75-CX0012
RTI Project Number 0217156
Award Amount: $1,197,273

January 1, 2020 – December 31, 2023

AI R&D to Support Community Supervision: Integrated Dynamic Risk Assessment for Community Supervision (IDRACS)

Final Report

Pamela K. Lattimore, Ph.D.
Principal Investigator
lattimore@rti.org
1.919.485.7759

Christopher Inkpen, Ph.D.
Project Director
cinkpen@rti.org
1.919.541.6411

RTI International [cover-address]
3040 E. Cornwallis Road, PO Box 12194 [cover-address]
Research Triangle Park, NC 27709 [cover-address]
www.rti.org

NIJ Grant Number 2019-75-CX0012
RTI Project Number 0217156
Award Amount: $1,197,273

The development of the Integrated Dynamic Risk Assessment for Community Supervision (IDRACS) tool was funded by Grant 2019-75-CX-0012 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

Acknowledgments

The evaluation team extends our many thanks to those who contributed to this evaluation. In particular, the team thanks the administration and staff of the Georgia Department of Community Supervision (DCS), who provided invaluable and collaborative support for the project; and the Georgia Department of Community Supervision and Georgia Crime Information Center, which provided data for the project. Special thanks are extended at DCS to Nick Powell and his research team as well as to the many individuals in the information technology group who provided invaluable assistance in the interpretation of DCS data, policy, and practices as the IDRACS tool was developed and integrated into DCS's case management system. The help and support of John Spier, Sharon Johnson, and their colleagues at Applied Research Services, Inc., with data and collaboration with DCS, is also acknowledged.

Authors

Pamela K. Lattimore, Christopher Inkpen, Stephen Tueller, and Caroline Kery, RTI International

Report Contributors

Kim Janda, RTI International; Sharon Johnson, Applied Research Services, Inc.; Luke Muentner, RTI International; Nicholas Powell, Georgia Department of Community Supervision; John Speir, Applied Research Services, Inc.

Contents

Figures

Number **Page**

Tables

Number **Page**

Executive Summary

This report describes the findings from the Integrated Dynamic Risk Assessment for Community Supervision (IDRACS) project, a research study led by RTI International in collaboration with Applied Research Services, Inc. (ARS) and the Georgia Department of Community Supervision (DCS) and funded by the National Institute of Justice by Award Number 2019-75-CX-0012 under the solicitation for Artificial Intelligence Research and Development to Support Community Supervision, FY2019.

Purpose

The purpose of this project was to develop accurate and responsive risk algorithms for the Georgia DCS. DCS is a statewide agency tasked with administering felony probation and parole, with an active supervised population of over 200,000 people. In addition to supervising men and women on parole, this population includes people sentenced directly to a probation term (i.e., straight probation) and those sentenced to incarceration followed by a term of community supervision (i.e., split probation). The goal of this study was to develop models for felony or violent rearrest (or revocation) that (1) improved upon current risk algorithms implemented in DCS's case management system, and (2) incorporated dynamic features that allowed risk scores to increase or decrease depending on an individual's progress on supervision.

Data and Methods

To develop risk algorithms for DCS, we used detail from before and during supervision for roughly 146,000 unique individuals who started felony probation or parole in Georgia between 2016 and 2019. Administrative data and their primary sources are:

(1) information gathered before and during supervision from DCS, (2) detail on prior

prison admission and relevant prison terms for those on parole from the Georgia Department of Corrections (GDC), and (3) criminal history on arrests and convictions from the Georgia Crime Information Center (GCIC). As a result of data cleaning and merging, RTI project staff produced a dataset of nearly 151,000 longitudinal supervision histories that encompass the period from the start of a supervision term until the last observed entry for that term. Longitudinal records in this dataset end with either the completion of a supervision term, the individual experiencing the study outcome of felony or misdemeanor violent arrest or revocation, or the end of the study observation period (12/31/2019). The data include detailed criminal history data and contextual information derived from the probation dockets or parole cases, such as the underlying charge or special conditions assigned during supervision. Furthermore, in addition to static factors such as conviction offense, the data include dynamic measures that change over the course of supervision, such as the number of positive or negative drug tests, violations, and misdemeanor arrests; changes in employment verification or supervision level; and presence of outstanding warrants.

To produce the risk algorithms, we explored a variety of approaches and methods. The project team assessed both a classification and survival (i.e., time to event) approach. Furthermore, these analyses included comparisons of traditional inferential statistical models (e.g., logistic regression models for classification, parametric or semiparametric models for survival analyses) and machine learning methods. The datasets were split to produce a training dataset for model development and test and validation datasets to assess model accuracy on held-out data. Models were compared based on model accuracy metrics, including area under the curve for classification models and

concordance and Brier scores for survival models. In addition to developing accurate models, this project involved a validation effort using supervision episodes that started after the widespread availability of vaccines related to the coronavirus disease 2019 (COVID-19) pandemic to ensure that models developed before the pandemic would be useful in post-pandemic settings.

Results

In accordance with DCS's current supervision practices, we produced separate models stratified by biological sex and supervision type (straight or split probation and parole). The model exploration process included:

- Comparing the utility of including detail on criminal history by arrest type and timing before supervision start
- Including dynamic features collected during the course of supervision that allowed for increases and decreases in an individual's risk profile
- Assessing the utility of developing period-specific logistic regression models that aligned with DCS's internal supervision practices for the first 90 days, the remaining three quarters of the first year, and after the first year of supervision.

The project team found that including detail on the nature and timing of the underlying criminal history produced more accurate results compared to models that used broad lifetime criminal histories. Furthermore, applying feature selection algorithms suggested that omitting arrests that occurred 5 years before the start of supervision did not worsen model accuracy. Tests of including dynamic measures revealed substantial gains in model accuracy. Additionally, period-specific models (compared to modeling

rearrest/revocation comprehensively) proved to be most accurate for predictions in the first year of supervision. Applying machine learning techniques revealed that while these models sometimes produced modest improvements in accuracy, they were often not significantly or substantively different in contrast to the tradeoffs in model interpretation and ease of implementation when compared to traditional statistical models (i.e., logistic regression). Net of producing accurate models, the RTI team also developed and implemented a process that entailed bootstrapping predictions to create confidence intervals around individual predictions, incorporating uncertainty into one's predicted probability of rearrest.

After providing details on model development, this report describes the process of integrating the IDRACS risk algorithms into DCS's case management system. Integration included providing documentation and syntax as well as carrying out a data management and prediction task on a shared dataset to ensure equivalent inputs and outputs. Lastly, the analysis efforts included a validation task that compared model accuracy metrics derived from the original cohort data compared to data collected for probation and parole starts after the COVID-19 pandemic. The results for the first quarter models (where similar observation periods were available) indicate that model accuracy for more recent supervision terms does not differ compared to the accuracy metrics for data on which the original models were developed.

Conclusions

The outcome of this project was a suite of predictive algorithms and data management processes that will supplement DCS's supervision practices, allowing for accurate and time-specific predictions of the risk of felony or violent rearrest or revocation. The utility

of these tools was vastly improved through extensive collaboration between the research team and DCS's operational, research, and information technology staff, as well as through concerted efforts to understand officers' interpretations of risk and the use of risk scores while on supervision. To complement community supervision officers' requests to reflect success on probation or parole, systematic efforts should be implemented to collect data that reflect both signs of reengaging in and desisting from criminal activity.

1. Introduction

The United States imprisons more people than any other country in the world (Travis et al., 2014). In 2021, 1.2 million people were incarcerated in the nation's prisons with another 636,000 confined in local jails (Carson, 2021; Zeng, 2022). However, the largest share of those under correctional oversight are those on community supervision (Travis et al., 2014). In 2021, more than 3.7 million adults were on supervision, including 2.9 million of whom were on probation and 800,000 on parole (Kaeble, 2023). When people enter the criminal legal system, and as they continue to navigate its various arms, they are often assessed for "risk" to inform care placements and develop intervention responses (Latessa & Lovins, 2010). Indeed, these risk assessments can be used to help criminal legal officials in the determination of security classifications, sentence planning, parole and release decision-making, treatment and rehabilitation, and even the management of special populations (e.g., juveniles, those with severe and persistent mental health issues, or individuals with a history of violent behaviors). However, the nature of community supervision presents a unique problem when developing risk assessments for the field. Specifically, individuals under supervision are in the community posing a risk to public safety; they are supervised at different levels of intensity; and the amount and nature of data that are collected varies over time.

This report describes the development and implementation of the Integrated Dynamic Risk Assessment for Community Supervision (IDRACS) tool, developed by RTI International in collaboration with the Georgia Department of Community Supervision (DCS) and Applied Research Services, Inc. (ARS). The goal of this project was to produce an accurate and useful tool for predicting felony or violent misdemeanor

rearrest (or revocation) for individuals on community supervision (both probation and parole) that improved upon the current risk assessment tool incorporated into DCS's case management system (CMS). Herein, we describe the process of developing a longitudinal dataset for assessing the likelihood of rearrest or revocation over a substantial time period for individuals on supervision in Georgia; testing multiple analytic approaches to estimation; validating the final set of models; and working with DCS to successfully transfer the models to their CMS. The final set of models incorporates dynamic as well as static risk and protective factors. Consistent with the previous models used by DCS, separate models were estimated for men and women and for three different types of supervision: probation, split probation, and parole. Further, to increase model accuracy, separate models were estimated for three periods of supervision corresponding loosely to supervision practices in Georgia: first 3 months (Period 1), next 9 months (Period 2), and one year or greater (Period 3).

This report first describes the IDRACS project and the context of felony probation and parole in the state of Georgia, including background on community supervision and the use of risk assessments in supervision settings. Chapter 2 provides a description of the data sources used to develop the longitudinal cohort, including details on the inclusion and exclusion criteria for observations followed by a description of the outcome and predictor measures. In Chapter 3, the report discusses the analytical strategies employed to develop and test models and test and compare model fit and accuracy. Chapter 4 describes the IDRACS cohort, introducing details on outcome measures and predictor variables by supervision type and sex. Chapters 5 and 6 present the results of the IDRACS models along with the results of a series of statistical tests that aided in the

development of the final models, including tests of model specification, model type, and modeling time periods. We then include, in Chapters 7 through 8, a description of the model integration process undertaken with DCS and an assessment of model performance during the COVID-19 pandemic and beyond. Chapter 9 provides limitations, conclusions, and recommendations of the study.

1.1 The Georgia Department of Community Supervision

Felony probation and parole in the state of Georgia is administered by DCS. DCS oversees felony probation for roughly 200,000 people in 50 judicial circuits in Georgia. Although felony probation is a sentence to prison, terms of straight probation are served in the community directly after sentencing, whereas split probation includes a sentence of incarceration to prison followed by a term to be served in the community. In contrast, parole is a conditional early release from a prison sentence to be served in the community. Including misdemeanor supervision, Georgia has the largest estimated population serving their sentences in the community, with roughly 370,000 people on felony or misdemeanor community supervision during the year. This number equates to roughly 1 in 23 adults in Georgia being on some type of supervision (Kaeble, 2021).

To address the issues of overcrowding in prison and the large, supervised population, in 2013 the Georgia general assembly passed legislation that created the Georgia Council on Criminal Justice Reform. As part of their initial analyses on criminal justice issues in the state, they identified that a unified statewide agency tasked with overseeing felony community supervision could benefit the state. Consequently, in 2015, the Georgia General Assembly passed House Bill 310 (HB310), which transferred responsibility for overseeing parole from the State Board of Pardons and Parole to DCS, along with

shifting felony probation from the Georgia Department of Corrections (GDC) to DCS, effectively creating the statewide agency tasked with comprehensive felony community supervision.

In 2023, DCS had more than 1,700 staff (including sworn and non-sworn personnel) working in 61 field offices spread throughout the 50 judicial circuits. Community Supervision Officers (CSOs, i.e., combined probation and parole officers) are required to have a 4-year college degree, and all officers complete a standardized basic training certification process known as the Basic Community Supervision Officer Training (BCSOT). CSOs in Georgia are *sworn*, indicating that they have the authority to enforce Georgia laws. In addition to sworn officers, DCS employs nearly 500 non-sworn personnel, which aid in the administration of community supervision throughout the state. According to DCS's internal estimates, officers have an average of 110 individuals on their caseloads. In addition to standard supervision caseloads, officers may have individuals on a "high" supervision level (which requires additional check-ins and scrutiny) or a "contact" supervision level (which requires telephonic check-ins). Furthermore, DCS carries out "specialized" caseloads for individuals who require additional supervision (e.g., sex offenders).

In addition to creating the statewide agency, HB310 also mandated DCS to used evidence-based practices for operations and supervision strategies. For example, DCS currently uses a risk assessment developed and validated by ARS, which uses stratified algorithms for men and women for individuals on straight probation, split probation, and parole. These algorithms are incorporated into DCS's CMS to produce risk scores, which are in turn used to set supervision levels. In addition, in accordance with HB310,

DCS has conducted internal evaluations of their Day Reporting Centers (DRCs). As such, the agency is well-versed in data collection, using research to inform their operations, and collaborating with research partners.

1.2 Risk Assessments in The Criminal Justice Legal System

The practice of measuring risk, defined as the likelihood of reoffending or not complying with legal requirements, has a nearly century-old history in the United States (see Bureau of Justice Assistance, n.d.). Early assessments relied on clinician or correctional staff judgments but were inevitably prone to human error and biases (Harcourt, 2015). This led to the need for more systematic, objective, and evidence-based risk assessment strategies. In the 1920s, the first tool using numeric predictions of stable risk factors was developed and introduced for the Illinois parole system (Bonta & Wormith, 2007; Connolly, 2003). The tool considered marital status, criminal and employment history, and institutional misconduct as static risk factors (meaning characteristics that are unchanging) to predict reoffending (Bonta, 1996). Although these assessments proved more reliable than personal judgments alone, they did not account for dynamic changes in attitudes, behaviors, and needs over time (Raynor, 2016).

In the 1980s, updated risk assessments incorporating variable characteristics related to reoffending, such as substance use and antisocial behaviors, emerged (Harcourt, 2015; Raynor, 2016). The risk-needs-responsivity (RNR) framework evolved from there, emphasizing tailored rehabilitation efforts based on an individual's risk level, criminogenic needs, and learning style and motivations (Andrews et al., 2011; Bonta & Andrews, 2007; Ogloff & Davis, 2004). The major risk/need factors measured in RNR

models include antisocial personality patterns, attitudes toward criminality, social supports for crime, substance abuse, family/marital relationships, education and employment history, and prosocial recreational activities; more minor needs that are assessed include self-esteem, personal distress, and mental and physical health (Bonta & Andrews, 2007). Not only do RNR models set out to predict risk, but they also strive to identify which criminogenic needs to target for intervention as a means of reducing future legal involvement and improving public safety.

Today, many risk assessments in the criminal legal system have evolved to integrate case management efforts and maximize the benefits of treatment and supervision. Several modern instruments demonstrate this approach, such as the Level of Service/Case Management Inventory (LS/CMI), Violence Risk Scale, Correctional Assessment and Intervention System, Public Safety Assessment (PSA), and Correctional Offender Management Profile for Alternative Sanctions (COMPAS). These tools encompass wide-ranging factors to assess risk and address individual needs across various phases of the criminal legal process. For example, the PSA assesses nine factors related to personal demographics (age) and prior court/legal involvement (Advancing Pretrial Policy & Research (APPR), n.d.). It has proven highly predictive in estimating pretrial failure to appear in court and new arrests while on pretrial release (DeMichele et al., 2020; Milgram et al., 2014), making the PSA widely used in court pretrial settings. Alternatively, the COMPAS assesses criminal involvement, relationships, lifestyle, personality and attitudes, and social exclusion across 22 different scales (Blomberg et al., 2010; Brennan et al., 2008). The various risk and need scales are designed to inform individual case management as it relates to supervision and

programming decisions (Blomberg et al., 2010; Brennan et al., 2008). The tool has become widespread in decision-making across the legal system, including in pretrial release, security placements, institutional programming, release, and community supervision.

1.2.1 Risk Assessments in Community Supervision

Risk assessments play a particularly crucial role in community corrections (i.e., probation and parole) (e.g., Burrell, 2016). Officials often use these measures to determine: (1) supervision level ranging from intensive supervision to regular check-ins; (2) reentry planning, including designing comprehensive plans that address housing, employment, family support, and other needs for successful reintegration; (3) case planning, including outlining individualized interventions that address underlying causes of criminality; (4) treatment allocation, such as directing resources to target criminogenic factors (e.g., substance use or mental health counseling); and (5) revocation decision-making, including informing decisions around continued community supervision or alternative sanctions. Several risk/need assessments have been used in community corrections settings to facilitate these processes, a few of which are summarized below.

The LS/CMI and the Ohio Risk Assessment System (ORAS) have both been widely implemented within probation and parole agencies (Andrews et al., 2000). Each is designed to be a comprehensive tool that assesses both static and dynamic risk, needs, and responsivity factors for adult populations. The 41-item LS/CMI combines risk assessment and case management into a single system, providing probation and parole officers with tools for holistic management and treatment planning (Andrews et al.,

2000). Although the LS/CMI has demonstrated strong predictive validity for recidivism within community corrections populations (Jimenez et al., 2018; Onifade et al., 2009), it has been critiqued for poor adherence of case management to RNR principles and heterogeneous findings across gender and racial groups (Dyck et al., 2018; Olver et al., 2013). The ORAS takes a different and more systematic approach, employing nine different tools throughout an individual's time in the legal system that use a number of measures or items to predict failure, including reentry (18 items), community screening (35 items), and community supervision (4 items) tools (Latessa et al., 2010). It, too, has demonstrated robust validity within community corrections (Latessa et al., 2010; Lovins et al., 2018), although concerns exist around its acceptability in new populations, given that it may not adequately address sociopolitical differences across settings (Lovins et al., 2018).

Other tools have been used within the juvenile community corrections space and among specific populations. For instance, the Positive Achievement Change Tool (PACT), which consists of 126 items over 12 domains, identifies both risk and strengths/protective factors in young persons' lives, thereby matching appropriate programs and services to individual needs (Baglivio, 2007). Although the PACT has proven effective in predicting juvenile recidivism (Baglivio & Jackowski, 2012; Winokur-Early et al., 2012), scholars have questioned the quality of its implementation across settings, and ongoing validation is necessitated (Mueller et al., 2022). Efforts being made in New York City reaffirm the need for context-specific, evidence-based risk assessment instruments (RAIs) that support juvenile reentry. Initial testing of the city's RAI holds promise for predicting court non-appearance and rearrest rates among youth

(Fratello et al., 2011); however, further comprehensive evaluation is necessary to fully assess its effectiveness.

Other tools have been developed and implemented that address specific types of offenders. For example, the Static-99R (ten items) evaluates risk for men convicted of sex offenses before release, focusing on static factors that predict the likelihood of sexual reoffending (Helmus et al., 2022; Phenix et al., 2016). It has demonstrated consistent predictive accuracy in terms of sexual recidivism (Brouillette-Alarie & Proulx, 2013; Hanson et al., 2016); however, variability is notable in the predicted rates across studies (Helmus et al., 2012; Helmus et al., 2022).

1.2.2 Accuracy And Bias in Risk Assessments

As with any assessment or prediction, risk assessment tools will not provide perfect prediction. Additionally, a standardized approach for identifying the likelihood for reoffending does not guarantee the elimination of bias in prediction. For instance, a recent critical examination of scores on the COMPAS assessment found that 20% of people who were predicted to commit violent crimes went on to do so, implying a false positive rate of 80%. Predictions of general reoffending were more accurate, with 61% of the people deemed likely to reoffend being arrested for crimes within 2 years (Angwin et al., 2016). Additionally, White individuals are significantly more likely to be categorized as low risk, while Black individuals are forecasted to reoffend at twice the rate of White individuals, leading to a higher false positive rate for Black individuals (Angwin et al., 2016). The COMPAS assessment is not the only assessment tool that has been critiqued for racial bias. Other work has discussed how risk assessments have higher predictive validity for White men compared to other populations, with factors such

as peer influence and criminal history having sporadic utility by jurisdiction and sample (Campbell et al., 2018; McCafferty, 2016; Vincent & Viljoen, 2020). Additionally, studies have shown that failing to account for differences in base rates between men and women can lead to overestimates of recidivism for females (Skeem et al., 2016). Despite these findings, very few validation studies of risk assessments conduct subgroup analyses by race or gender (Singh et al., 2013).

To assess model performance, we use the area under the receiver operating characteristic (ROC) curve. The ROC plots the true positive rate compared to the false positive rate at every threshold. From the ROC curve, the area under the curve (AUC) is calculated to provide a comprehensive, threshold-neutral statistic that represents the probability that the model correctly predicted the observed class (e.g., arrest) (Huang & Ling, 2005). AUC values can range from 0 to 1.0, with 0 being perfectly incorrect prediction, 0.5 akin to random chance (i.e., flipping a coin), and 1.0 being perfect prediction. Research evaluating criminal justice risk assessments identifies the AUC values associated with risk instruments as poor (0.5–0.54), fair (0.55–063), good (0.64–0.70), or excellent (0.71+) (Desmarais et al., 2017). We also assess the final models for racial bias.

2. Data Sources and Measures

Creating a legible and linear series of events from data drawn from the criminal legal system can be difficult, as practitioner data are traditionally collected for individual case management and rarely designed with research in mind. This chapter describes the data sources used to identify our study cohort and create longitudinal supervision histories for people who started either felony probation or parole in the state of Georgia between 2016 and 2019. The chapter starts by describing the data sources used in the study. The chapter then turns to a discussion of the steps required to create longitudinal supervision histories, the inclusion and exclusion criteria used to establish the cohort, and the CONSORT diagram that describes records that were filtered out in the process of finalizing the analytical cohort dataset. We then move to a discussion of the static and dynamic measures used in modeling, before closing with a description of the study outcome variable.

2.1 IDRACS Study Period and Analytical Dataset Design

Administrative data were obtained per executed data use agreements with three Georgia agencies: DCS, the Georgia Crime Information Center (GCIC), and GDC. The data include information on 348,577 unique individuals who were on community supervision in Georgia between January 1, 2013, and December 31, 2019. The cohort was subsequently reduced to include only the 160,428 individuals who *began* a supervision term on or after January 1, 2016. These data were then processed and merged to produce the final dataset that was randomly stratified to produce training/estimation (70%), validation (10%), and test (20%) datasets for model estimation and validation. Following existing practices at DCS, separate models were to

be developed for different types of supervision—straight probation, split probation, and parole—and for men and women. Thus, the goal for the data development was to produce six non-overlapping longitudinal datasets. Because individuals could be serving overlapping probation terms or probation and parole terms, constructing an appropriate dataset was challenging.

2.2 Data Sources

The Georgia DCS data were provided from DCS's supervision management database. The database was created in 2016 after DCS was formed in 2015 to merge probation and parole supervision under a single agency. The data, originally stored as a dynamic relational database, contain a variety of tables, including officer interactions, violations, program attendance and referrals, supervision terms and supervision-level changes, prison and custody terms, arrests, and probation dockets. These data were used to construct a complete history of each person's supervision. State identifiers for individuals in the DCS cohort were provided to the GCIC, which provided complete Georgia arrest history data for the individuals. Arrest information was also available in a DCS table.[1] The GDC data provided information on prison incarceration histories, including the dates of prison terms and a few other variables, such as gang affiliation while incarcerated.

2.3 Supervision Periods

An integral component of developing the analytical cohort dataset was to establish supervision periods for every individual. Although seemingly straightforward, this task

[1] Because DCS receives their arrest data from GCIC, these two datasets were similar, although the GCIC arrest data include some information on older arrests, and the DCS data include some out-of-state arrests when these were known by the supervising officer.

was complicated by the fact that individuals can have multiple probation dockets active at the same time and can be on both parole and probation. The supervision context data were provided in a set of nested tables that were used to create an individual's overarching ***supervision period***. ***Supervision "histories"*** covered a period of continuous supervision. Inside those supervision histories, an individual could have multiple ***supervision "episodes"*** of probation or parole. Probation terms were then mapped to court dockets to identify contextual information about the term, while parole terms were mapped to GDC case numbers. Given that individuals could have multiple supervision histories that often overlapped, the solution was to create "***mapped" supervision periods*** by taking all overlapping histories and merging them.[2] This was done to obtain the overall start/stop dates for parole and probation. These mapped supervision periods were then classified by whether they contained parole, probation, or both types of supervision episodes.

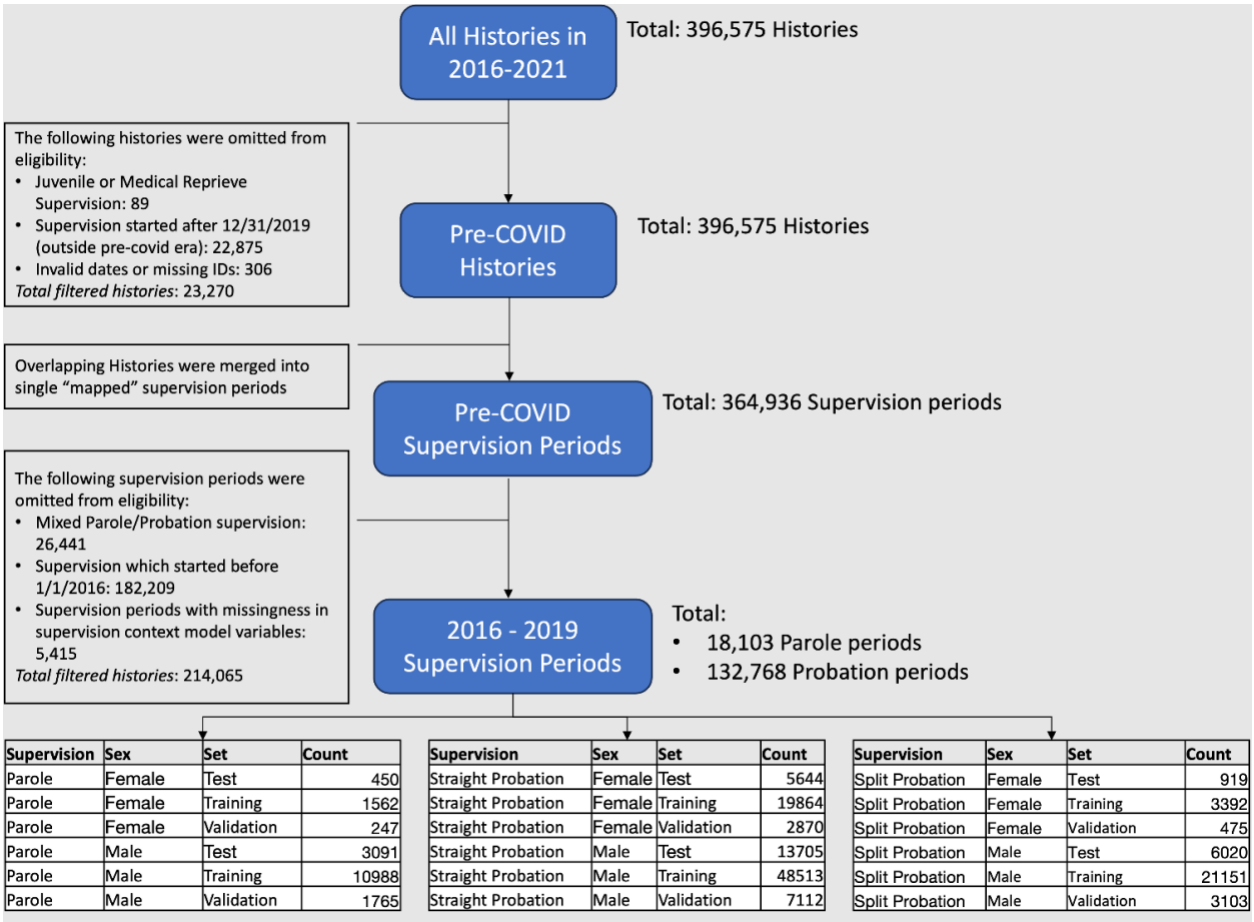
2.4 Exclusions

Once the mapped supervision periods were created, records were excluded for several reasons. Figure 2-1 shows that the original data included 396,575 supervision histories. This number was reduced through multiple exclusions to produce the final analytic dataset. The most significant impact is attributable to the decision to right- and left-truncate the data. As the DCS database was created in 2016, missing information was substantial for events that occurred for people on supervision before then. Therefore, the data were left-truncated to exclude anyone whose supervision history started before

[2] The data frequently had duplication issues, because DCS would create new records when changes needed to be made on an individual's supervision start or end date. For this reason, de-duplication was performed to identify a person's relevant supervision period.

January 1, 2016 (N = 182,209). Similarly, to avoid the impact of any policy and practice changes after the onset of the coronavirus disease 2019 (COVID-19) pandemic, the data were right-truncated as of December 31, 2019 (N = 22,875).

Figure 2-1. Development of the Estimation and Validation Data



Supervision	Sex	Set	Count
Parole	Female	Test	450
Parole	Female	Training	1562
Parole	Female	Validation	247
Parole	Male	Test	3091
Parole	Male	Training	10988
Parole	Male	Validation	1765

Supervision	Sex	Set	Count
Straight Probation	Female	Test	5644
Straight Probation	Female	Training	19864
Straight Probation	Female	Validation	2870
Straight Probation	Male	Test	13705
Straight Probation	Male	Training	48513
Straight Probation	Male	Validation	7112

Supervision	Sex	Set	Count
Split Probation	Female	Test	919
Split Probation	Female	Training	3392
Split Probation	Female	Validation	475
Split Probation	Male	Test	6020
Split Probation	Male	Training	21151
Split Probation	Male	Validation	3103

Additionally, some individuals had overlapping probation and parole sentences. For example, an individual had split sentence probation and was released from prison on parole, leading to an overlap of parole and probation sentences; or an individual was on probation and was subsequently sent to prison for another crime, and was then released on parole without the original probation term ending. Because of the complexities of deriving contextual information for mapped periods with both probation

and parole supervision types, we chose to drop mapped supervision terms with overlapping parole and probation supervision types (N = 26,441) from our cohort and focused instead on parole and probation periods individually. In addition, we excluded juvenile and medical reprieve supervision (N = 89) from the cohort.

The final dataset included 18,803 parole supervision periods and 132,768 probation supervision periods. For analysis purposes, this final dataset was stratified by supervision type (straight probation, split probation, parole) and sex (men, women). Each of these six datasets was then further divided into a test dataset (20%), training dataset (70%), and validation dataset (10%), used for model development and accuracy testing.

In addition to the analytic cohort datasets, additional data were processed to support analyses examining the validity of the developed models following the onset of COVID-19. These data covered the "COVID period," which was defined as March 2020 through June 2021 when the first COVID-19 vaccine became available, and the "post-COVID period," which was defined as July 2021 through October 2022. Results of these supplemental analyses are reported in Chapter 7.

The static and dynamic factors anticipated to be useful for model development, as well as the outcome variable (felony or violent misdemeanor arrest or revocation), are described next. Table 2-1 shows the measures, variable types, and range of values for factors that were derived to predict rearrest while on probation or parole. The measures here are described for the final specification of the model, but Chapters 6 and 7 provide detail on the decision-making processes and the data limitations that led to the inclusion of the following measures.

2.5 Static Factors

To account for an individual's history and factors associated with their probation and parole term, we include a series of measures derived at the supervision start date (Table 2-1). These static factors, which did not vary over the course of a supervision period, included demographic **characteristics, criminal history information, and contextual information about the supervision period.** Demographic information was limited to date of birth, which was used to calculate age at the start of supervision, gang affiliation, and race.[3]

Table 2-1. Static Measures for Probation and Parole Periods

Measure	Type	Range or Values
Age at supervision start (in years)	Numeric	18–99
Race	Categorical	0 = "White", 1 = "Non-White"
Confirmed gang membership	Categorical	0 = "No", 1 = "Yes"
Prior arrests (0–2 years and 0–5 years)		
Violent offenses	Numeric	0–10
Public order offenses	Numeric	0–21
Drug offenses	Numeric	0–20
Property offenses	Numeric	0–33
Probation/parole offenses	Numeric	0–16
Prior prison terms (0–5 years)	Numeric	0–3
Prior probation or parole (0–5 years)	Numeric	0 = "No", 1 = "Yes"
Most serious charge for probation docket	Categorical	Reference = Drug charge 1 = "Other charge" 2 = "Property charge" 3 = "Public order charge" 4 = "Violent charge" 5 = "Missing offense detail"
Parole-specific measures		
Underlying parole offense is a property charge	Categorical	0 = "No", 1 = "Yes"
Prior prison admission was for revocation	Categorical	0 = "No", 1 = "Yes"

[3] Race was used only to train the models, but was withheld when making predictions so that any racial bias present in the training data would be absorbed by the unused race coefficient. Race was used later when evaluating bias in model results.

Measure	Type	Range or Values
Flagged for mental health treatment in GDC	Categorical	0 = "No", 1 = "Yes"
Prior prison disciplinary reports	Numeric	0–99
Any special parole conditions	Categorical	0 = "No", 1 = "Yes"

Criminal history measures were past arrests, past prison terms, and past supervision terms. For arrests, we used the National Corrections Reporting Program (NCRP) charge categorization (Perkins, 1993) to assign arrest charges (determined from their GCIC crime code) to the broad categories of (1) violent, (2) property, (3) drug, (4) public order, and (5) probation and parole offenses. We then counted the number of unique arrests within each offense category as a numerical variable. Rather than considering lifetime arrest history, we developed a set of measures that provided limited lookback at the arrest history before the start of supervision (e.g., 0–1 years, 0–2 years, 0–5 years). These measures allowed us to determine whether better results could be obtained by focusing on more recent criminal behavior and, indeed, lookback periods of 2 or 5 years, in contrast to lifetime, provided best fits for the models. Prior prison terms were similarly defined within each of these time bands. Identifying past supervision terms was more complex because many supervision tables were not created until 2016. The tables that did exist provided information on supervision-level events, drug tests, and violations. Thus, the presence of events in any of those tables was used as evidence of past supervision, with the acknowledgment that pre-2016 this evidence was mostly for past parole.

Information on the probation term context was derived from the court docket, which led to the probation sentence. The dockets included information on offenses and the

sentencing that resulted from them. The dockets were used to categorize probation terms as “split” (the sentence was prison followed by probation) and “straight” (the sentence was probation only with no initial prison time served). Split status was derived by comparing the overall sentence to the time to serve fields (split would have a sentence longer than the time served) and the presence of a “probation start date” on a docket with time served. The offenses related to the sentence were, like arrests, converted into NCRP categories. The mid-level NCRP crime codes are numbered in order of severity, so we used the codes to find the most serious charge on each docket. Unlike the arrest tables, however, no convenient lookup was available to translate between the offense code in the docket table and the NCRP category. Instead, we made use of the Rapid Offense Text Autocoder (ROTA) tool (Baumgartner et al., 2021) and the offense descriptions to map the docket charge to their most likely NCRP category.

As shown in Table 2-1, we also incorporate information on the context of the parole case, which includes information derived from an individual’s time in prison and details on special conditions assigned to the parole case. In the existing model used by DCS, parole supervision predictions were informed by several prison-term-related fields that came from the GDC data.[4] The prison variables that were extracted for potential inclusion in the models flagged the identification of mental health treatment during prison, details on the crime that led to the prison episode (e.g., property crime or not), the number of disciplinary reports while the individual was in prison, and whether the

[4] The GDC parole term and the DCS data had no direct links, so the prison term that ended near the start of the parole term was assumed to be the corresponding prison term.

prison admission was for a revocation from supervision. In addition, we include a categorical measure that indicates if any special conditions are associated with the parole term.

2.6 Dynamic Factors

Factors that capture events or changes in status during supervision were also identified, including running "counter" variables and dichotomous "on/off" variables. In Table 2-2, these measures are listed as either numeric (counter) or categorical (dichotomous) variables. Counter variables measure the number of times an event happened since the start of supervision, while dichotomous variables identify if a particular event or status was ongoing at the time of prediction. These measures were categorized as being dynamic *protective* factors or dynamic *risk* factors, as their inclusion could reveal progress on supervision or an increased risk of experiencing the outcome.

Table 2-2. Dynamic Measures for Probation and Parole Periods

Measure	Type	Range or values
Drug Testing		
Positive drug tests (any positive test on a date)	Numeric	0–99
Negative drug tests (all tests listed as negative on a date)	Numeric	0–99
No drug testing in previous 90 days	Categorical	0 = "Yes testing" 1 = "No testing"
Employed	Categorical	0 = "Unemployed" 1 = "Employed"
Violations during supervision (any type)	Numeric	0–99
Active warrant	Categorical	0 = "No", 1 = "Yes"
Moved to contact status	Categorical	0 = "No", 1 = "Yes"
Count of supervision-level changes during time period	Numeric	0–99
Probation Conditions		

Measure	Type	Range or values
Community service	Categorical	0 = "No", 1 = "Yes"
Drug or alcohol restrictions	Categorical	0 = "No", 1 = "Yes"
Drug or alcohol testing	Categorical	0 = "No", 1 = "Yes"
Education	Categorical	0 = "No", 1 = "Yes"
Employment	Categorical	0 = "No", 1 = "Yes"
Fees	Categorical	0 = "No", 1 = "Yes"
No contact orders	Categorical	0 = "No", 1 = "Yes"
Other	Categorical	0 = "No", 1 = "Yes"
Violence-related conditions	Categorical	0 = "No", 1 = "Yes"
Misdemeanor Arrests During Supervision		
Drug offenses	Numeric	0–99
Public order offenses	Numeric	0–99
Property offenses	Numeric	0–99
Probation/parole offenses	Numeric	0–99

Counter variables used in the final model(s) included the number of technical violations, number of positive drug tests, number of negative drug tests, number of technical violations, supervision-level changes (e.g., going from a “high” to “standard” supervision level), and running counts of non-violent misdemeanor arrests accumulated during supervision assigned to NCRP broad categories (i.e., drug, public order, property, or probation/parole offenses). Other counters, such as the number of programs attended, the number of times someone changed residences, and the number of probation or parole delinquent reports filed, were explored but ultimately found not to be viable for prediction.

In addition to the running counter variables, we also developed dichotomous dynamic measures, which would turn on or off depending on if the status or event was observed during supervision. For example, employment verification was used to update a dichotomous *employment* variable that recorded whether someone was employed (part-

or full-time) or not. Similarly, on DCS's recommendation, we developed a categorical measure that indicated if the individual on supervision had been drug tested during the previous 90 days, which accounted for the selective nature ("testing for cause") of drug testing while on supervision after the initial intake period.

Other dichotomous dynamic variables captured probation and parole conditions. Probation conditions were identified from the "special conditions" field in probation dockets. Unlike the most severe offenses or split status, which were determined by the dockets that were open at the start of probation, probation conditions could turn on and off as different dockets began and ended within a probation term. Probation conditions were identified using lists of keywords that indicated a particular kind of condition was activated (see Appendix A). Special conditions on probation include community supervision, drug or alcohol prohibitions, drug or alcohol treatment, continuing education, employment, fines and fees, non-contact orders, violence-related conditions (e.g., anger management classes), and other conditions. For example, drug treatment-related conditions were identified through the presence of terms such as "drc" or "rsat," which indicate that the individual had to report to a *day reporting center* or *residential substance abuse treatment* center. These term lists were compiled through careful review of docket condition examples and were confirmed with DCS. Parole conditions were simpler, as DCS collected a table of parole case numbers tied to a small number of condition categories (e.g., mental health counseling and vocational training). For analysis, we binned the parole condition categories to line up with the probation condition categories. Parole conditions were treated as dichotomous; although the conditions were "on" through the entire time the person was on parole, they could be

treated as static for pure parole terms. Along with conditions, DCS also had a “needs” table, collected at intake, which recorded special needs such as homelessness or needing employment. However, because of a high rate of missingness, the needs events could not be used in the final models.

Other dichotomous variables reflect events that may have occurred while the person was on supervision, or the type of supervision they were on. These events include prison terms, custody events (including local jail custody, U.S. Immigration and Customs Enforcement [ICE] custody, out-of-state supervision), and supervision level. Warrant supervision indicates a warrant is out for the person’s arrest, and a contact supervision status requires less frequent verifications conducted by telephone. The custody events—specifically the automatically generated arrest custodies—were used in the integration stage to identify supervision restarts after arrest (see Chapter 6).

Finally, we examined indicators for recorded residential status and address changes, as well as employment indicators. These measures were explored for potential inclusion in the analytical models but were omitted after discussions with DCS that data collection practices for these measures were varied over the course of supervision. For example, employment verification is standard practice during the first year, but after one year is conducted only depending on progress on supervision. After discussion with DCS, a determination was made that residence entries reflected address verification, and individuals with verified entries were those who were able to be verified. This exploration highlights the importance of collaboration with the research partner during data management and modeling, because the data generation process may change with time.

2.7 Outcomes

The risk assessment outcome is serious arrest—defined as a felony or violent misdemeanor arrest—or revocation, whichever occurred first. This outcome was defined after consultation with individuals at Georgia DCS, who indicated that non-violent misdemeanor arrests did not necessarily result in a change in supervision practices or revocation. By defining the outcome in this way, we can include non-violent misdemeanor arrests that occur during supervision as predictor measures. A violent arrest was any arrest that included a violent charge based on the NCRP classification of violent charges. As shown in Table 2-3, violent charges included aggravated assault, arson, armed robbery, and rape. Revocations were defined differently for those on probation versus parole because of the different actions taken in response. Based on review of the data, a parole revocation led to a subsequent ending of parole, so for individuals on parole, a single revocation was the outcome. For probation, the data showed that an individual could experience a revocation of one of the dockets attached to the supervision episode and remain on probation. The revocation outcome for probation was therefore defined as occurring when all dockets attached to the probation episode were revoked.

Table 2-3. Charge Descriptions from National Corrections Reporting Program for Violent Charges

Aggravated assault
Murder
Simple assault
Hit and run driving
Violent offenses - other
Child abuse
Armed robbery
Blackmail/extortion/intimidation
Forcible sodomy
Lewd act with children
Rape - force
Sexual assault - other
Rape - statutory - no force
Manslaughter - non-vehicular
Manslaughter - vehicular
Unspecified homicide
Voluntary/nonnegligent manslaughter
Kidnapping
Assaulting public officer
Note: Free text descriptions from GCIC definitions of charges were classified to the NCRP mid-level charge categories.

2.8 Time Periods for Arrest

Exploratory analyses of the occurrence of serious arrests revealed unique patterns depending on the time period under examination. Specifically, arrests in the first 90 days were more likely to feature a new criminal charge compared to a felony arrest associated with probation or parole violation. This period also coincides with the initial intake supervision period for those entering supervision in Georgia. In contrast, for those who progressed in supervision past one year without a serious arrest, the most likely arrest among those who experienced an arrest was for a felony probation or parole violation. Thus, we explored developing discrete models for three time periods:

(1) the first quarter of supervision (Period 1), (2) the remaining three quarters of the first year (Period 2), and (3) one-plus years on supervision (Period 3). This temporal stratification is in addition to the previously described stratifications on supervision type and sex. Models were estimated using data for those who entered the risk period. Specifically, all individuals were included in the first 90-day models, while only those who were successful (no serious arrest or revocation) during the initial period were included in the second period. Correspondingly, only those who were successful through the first year of supervision were included in model development for the third period. The analytical framework for testing the benefit of modeling arrest in unique time periods is discussed in the next chapter.

3. Analytic Strategy and Methods

This chapter presents the analytical strategy used to develop the models predicting serious arrest (felony or violent misdemeanor arrest) or revocation for individuals on felony probation or parole in Georgia. While this project was driven by research questions centered on identifying the most accurate and viable models, the process of identifying useful predictors, model specifications, and model types was an exploratory and collaborative process with the Georgia DCS. Activities included testing different types of models, both in approach (i.e., survival vs. classification) and type (e.g., logistic regression versus machine learning [ML] classification models). In addition, as described in Chapter 5, we tested different model specifications or combinations of predictor variables. As described in Chapter 2, these measures were either static, and measured at intake, or dynamic, and measured over the course of supervision. In all analyses, the outputs of the models are compared for predictive accuracy.

This chapter starts by describing the modeling approach, which involved setting up the data to provide useful measures of predictive accuracy and developing and comparing classification and survival models. We describe the measures of predictive accuracy used in both the survival and classification approaches as well as the different types of survival and classification models used when identifying the most accurate and useful model types. In addition, this chapter describes the feature selection algorithms that were used to assist in model specification.

3.1 Approach

To produce a set of accurate and actionable models of serious arrest of the men and women being supervised by DCS, we employed several methods and model types to

identify the most accurate and useful models. This chapter describes the way in which we used data management and statistical testing to assess a model's predictive accuracy. We first describe the data management practices of splitting our cohort up randomly into training and test datasets, which allows for an assessment of a model's predictive accuracy while reducing the risk of model overfitting. We then turn to a discussion of different modeling perspectives and subsequent model types. Lastly, we discuss methods for assessing predictive accuracy between model types, model specifications, and in different subpopulations or time periods.

3.2 Training/Test Dataset Generation

The data were stratified on supervision type and sex and were then split into training or estimation, test, and validation datasets. This strategy is recommended in predictive modeling when sufficient data are available to develop ("train") a model, compare model performance using validation data, and finally assess its prediction, or generalization, error by predicting on the remaining held-out test data (Hastie et al., 2001). This practice avoids developing a model that overfits the data it was trained on, while allowing for estimation of performance in a production setting when the model will be presented with data not seen previously. The use of a separate validation dataset allowed us to assess and compare the model performance of varying model types and variable sets while continuing to hold out the test set for estimating the performance of the final model. For our model development, 70% of the data were used as training data, 10% as validation data, and 20% as test data.

3.3 Modeling Arrest in a Longitudinal Cohort

Predicting recidivism or serious arrest/revocation can be addressed as either a classification or a failure or survival problem. In classification problems, a dichotomous outcome measure (e.g., arrested or not) is observed at one point in time, and multiple model types or model specifications can be employed to best predict the outcome (Hastie et al., 2001). Survival (or time to event) analysis focuses on examining factors associated with differences in the timing of an event, as opposed to just the occurrence of the event (Altman & Bland, 1998). This study explored both survival and classification methods to predict arrest/revocation during community supervision. The methods tested for prediction are described next.

3.3.1 A Survival Approach to Modeling Arrest/Revocation

Recidivism, which is in this study considered a serious (felony or violent misdemeanor) arrest or revocation while on community supervision, is a classic survival analysis question and has been used in many analyses to explore the timing of arrest, including identifying the predictors associated with arrest and assessing the impact of interventions (Chung et al., 1991). Survival models are particularly beneficial when individuals are observed for different periods of time, as is the case with our data, because the full periods of observation can be used for all individuals—unlike in classification models, where observation must be truncated at the minimum time observed in the data.

This study examines the factors associated with increases or decreases in the timing of recidivism over a maximum 4-year study period, and compares models based on metrics of accuracy (described below). As a baseline, we use the Cox semiparametric

survival model to estimate the association between factors and the timing of serious arrest/revocation (Kleinbaum & Klein, 2012). Cox models are useful not only in assessing group differences or the impact of baseline characteristics in differences in time to events, but can also be used to incorporate time-varying measures (Zhang et al., 2018).

In addition to exploring the use of traditional inferential survival models, we also test the performance of several ML survival techniques, including tree-based methods (random survival forests and gradient-boosted trees) and neural network-based methods (deep survival analysis). Random survival forests are applied similarly to the random forest classification algorithm, wherein multiple decision trees trained on the longitudinal time event data are averaged to produce a prediction (Ishwaran et al., 2008). Gradient-boosted machines (GBMs) are tree-based models developed similarly to random forests; however, a GBM trains trees sequentially as opposed to all at once (as in random forest) with the goal of improving subsequent tree predictions (Chen et al., 2013). Lastly, we examine the utility of deep learning methods for survival analysis, which adapt multilayer probability models to a survival framework (Ranganath et al., 2016). We use two different specifications of deep learning methods to produce survival predictions: DeepSurv and RNN-SURV (Giunchiglia et al., 2018; Katzman et al., 2016). This wide variety of survival model types can be assessed for relative performance by using metrics that judge the model's accuracy, as described below.

The evaluation of the predictive accuracy of survival models necessarily must include the component of time in the assessment, as opposed to simply judging accuracy in overall predictions, as is done in classification problems. This study employs two

metrics (concordance and Brier score) to measure a survival model's ability to accurately predict the likelihood of being arrested or revoked.

Concordance can be used as a measure of model accuracy in a survival framework (Kremers, 2007; Penciana & D'Agostino, 2004). Random pairs of observations are selected, and an assessment is made on how the model orders the pairs based on their baseline characteristics. The concordance is then the fraction of pairs that are correctly ordered. Like the area under the ROC curve (hereafter AUC, or area under the curve) value (described below), concordance can be evaluated on a scale of 0.5 to 1.0, with 0.5 being equivalent to random chance and 1.0 indicating perfect prediction. The Brier score is another metric for assessing the accuracy of a survival model. The Brier score is the average squared distance between an individual's observed status (e.g., arrested or not, represented as a 1 or 0) and the predicted survival probability (between 0 and 1) (Park et al., 2021). Both measures are useful to assess how accurately a survival model predicts the timing and occurrence of an event in a longitudinal dataset.

3.3.2 A Classification Approach to Modeling Arrest/Revocation

Although the data in this analysis are longitudinal in nature, classification (or predicting a dichotomous outcome) is still a useful method for developing and implementing predictions of rearrest. Traditional recidivism analyses using a classification approach are conducted by identifying the minimum time of observation across the individuals in the data and taking the values for each individual at that time. Thus, for example, if everyone is observed for at least one year but some individuals are observed for only one year, the classification model will be restricted to a one-year period. Under this

example, an individual who experienced a serious arrest on one year and one day would be classified as not recidivating for the model.

To incorporate some elements of dynamic modeling into the classification models, as noted at the end of the previous chapter, we develop period-specific models that seek to simulate the approach of a survival model, known as survival stacking (Craig et al., 2021). This approach allows for including dynamic measures that can change value between the time periods, as dichotomous measures can change and running count measures can be incremented. This analysis employs logistic regression models to gauge the utility of different model specifications and the value of modeling rearrest in different time periods, as opposed to using the last observation in the longitudinal dataset. The process of model evaluation is described below.

3.3.3 Machine Learning Classification

In addition to identifying the most accurate and useful model specification, this project used ML classification methods to assess whether these methods yielded significant and substantive improvements in performance compared to traditional logistic regression classification models. There is no shortage of ML methods that could be used to predict rearrest, each with strengths and weaknesses. For our analyses, we employed tree-based methods because (1) tree-based methods, such as random forest, have been used for recidivism prediction (Berk et al., 2006, 2016; Berk & Bleich, 2013); and (2) as tree-based methods, such as random forest or gradient-boosted trees (i.e., XGBoost), are extensions of a decision tree, interpretability is straightforward in applying the method compared to other ML methods. Results of these methods were compared against the findings of a logistic regression classification algorithm.

Decision trees in general and their extension of random forests are useful in prediction, especially when tasked with identifying nonlinear relationships and local interactions between predictor variables (Bou-Hamad et al., 2011). Where decision trees identify variables and cut points that yield the most discrimination between classes, random forests are an ensemble method that pools and averages the predictions of multiple decision trees trained on different subsets of the data and using a selection of predictors (Speiser et al., 2019). Gradient-boosted trees offer an extension on this ensemble method that pools decision tree votes. As opposed to equal votes regardless of the selected tree, GBMs use sequential learning driven by a loss function that seeks to ensure that later trees are more accurate compared to earlier iterations (Natekin & Knoll, 2013). In addition to testing different ML classifiers, we also assessed the impact of handling class imbalance (i.e., arrest as a rare event) by under- and over-sampling (Gosain & Sardana, 2017).

3.3.4 Assessing Model Performance

To assess model performance, we use the area under the ROC curve. The ROC curve plots the true positive rate compared to the false positive rate at every threshold. The true positive rate is defined as the proportion of those who experienced the observed class (e.g., arrested) with a predicted probability above the threshold, and the false positive rate is the proportion of those who did not experience the observed class with a predicted probability above the threshold. From the ROC curve, the AUC is calculated to provide a comprehensive, threshold-neutral statistic that represents the probability that the model prediction correctly predicted the observed class (Huang & Ling, 2005). AUC values can range from 0 to 1.0, with 0 being perfectly incorrect prediction, 0.5 akin

to random chance (e.g., flipping a coin), and 1.0 being perfect prediction. Research evaluating criminal justice risk assessments identifies the AUC values associated with risk instruments as poor (0.50–0.54), fair (0.55–0.63), good (0.64–0.70), or excellent (0.71+) (Desmarais et al., 2017).

To gauge improvements in model performance when using different model types or specifications, we start by calculating AUC values derived from the test datasets. We then compare the predictive performance of different models by running significance tests on the differences between two AUC values using the pROC package in the R statistical computing environment (Robin et al., 2011). Model type performance can be gauged by examining how two different model types perform on the same dataset with the same model specification (i.e., different classifiers but the same predictors) and using a paired test of differences in AUC values. Model specification is assessed by running two different model specifications (i.e., different variables in the righthand side of the question) on the same dataset and using a paired test. To assess differences in models with different sample sizes, we use unpaired tests, such as when we compare the performance of a model on different subpopulations (e.g., White and non-White individuals). We also use paired tests when comparing models built using the same datasets (e.g., comparing static variables only vs. static and dynamic measures). Specifically, in Chapter 5, we use tests of differences in AUC values to compare (1) model types (e.g., logistic regression compared to random forest models), (2) model specifications (e.g., models using only static variables vs. models with static and dynamic measures), (3) modeling time periods (e.g., using one dataset vs. time period-specific datasets), and lastly, (4) differences in predictive accuracy by race. The

statistical tests from the pROC package give the probability that the observed difference in accuracy between the two models is due to chance. However, given the large sample sizes for specific subpopulations in our data (e.g., men on straight probation), even minor differences may be significantly different at “typical” 0.05 levels. Thus, in addition to identifying statistically significant differences and requiring that differences be significant at the $p < 0.001$ level, we assess if different model types or model specifications provide substantive improvements in accuracy by increasing the AUC by a certain value (e.g., between 0.03 and 0.08 absolute increase in the AUC).

3.3.5 LASSO Regression for Variable Selection

In addition to assessing differences in model types, we also explored several different model specifications (i.e., sets of predictor variables). DCS collects an expansive number of measures on the individuals on felony probation or parole. In addition to reviewing model output and comparing predictive accuracy, we also used feature selection algorithms to assist in assessing specifications of variables as well as identifying the most parsimonious yet accurate models. Maximizing model parsimony (i.e., the process of identifying the fewest number of predictor measures that can accurately predict an outcome) is useful, as it can aid in the development of models that are less likely to overfit the data on which they are built.

To identify a parsimonious set of predictors to predict felony or violent rearrest on supervision, we use the least absolute shrinkage and selection operator (LASSO) regression model (Muthukrishnan & Rohini, 2017). The predictor variables in this context are known as features, and LASSO is a method of feature selection. The LASSO regression method selects variables by forcing the regression coefficients of

weaker predictors toward zero; this is referred to as shrinkage and is implemented by adding a penalty parameter lambda to usual least squares regression. Lambda is selected by the data analyst and can be thought of as a “budget” for the sum of the absolute value of the regression parameters. A smaller value for lambda will shrink more regression coefficients toward zero to stay “within budget,” while a large value of lambda will allow more coefficients to be nonzero. In other words, lambda controls the number of features that are retained when implementing LASSO regression. For this study, LASSO regression was implemented using the glmnet (Hastie & Stanford, 2016) R package (Baldwin et al., 2010a), and lambda was selected using tenfold cross-validation for evaluating a range of possible lambdas.

3.4 Methods Summary

To summarize, our approach involved estimating a series of statistical models (survival models and logistic regression classification models) using a variety of factors derived from Georgia administrative data. In addition, ML methods were applied to the data to determine whether these techniques would yield superior prediction accuracy. Model specification with respect to factor inclusion was optimized using LASSO techniques. All models were fit on 70% of the data, validated against 20% samples, and tested on the remaining 10% sample.

Models were estimated for stratified subpopulations based on supervision type (straight probation, split probation, and parole) and sex (men and women). Based on exploratory analyses, an additional stratification by supervision time period (first quarter, next three quarters, one year or greater) was determined to improve model predictive accuracy.

Model fit, depending upon specification, and improvements in fit were assessed by concordance scores, Brier scores, or the AUC.

4. The IDRACS Cohort

This chapter provides descriptive detail about the longitudinal data used for the study, including descriptions of both predictor and outcome measures for the study cohort. We start by examining our analytical cohort, which includes all individuals placed on supervision from 2016 to 2019. We include individuals on straight or split probation, as well as those on parole. Given that our models are stratified by sex, we present descriptive statistics for male and female individuals. In addition, given that these analyses are longitudinal, and we employ an approach that divides the dataset by period, we present the distributions of outcomes and predictors by three time periods that were established through preliminary analyses. Period 1 is the first quarter of supervision, Period 2 is the remaining three quarters of the first year, and Period 3 is one year or greater of supervision. Individuals are included in each period unless they experienced a felony or violent misdemeanor arrest, their supervision was revoked, or their supervision otherwise ended (e.g., because of death or deportation). This chapter proceeds by discussing the outcome, defined as a serious arrest (felony or misdemeanor violent arrest) or revocation; the static measures captured at intake for community supervision; and dynamic features collected throughout the course of community supervision.

4.1 Outcome: Serious Arrest or Revocation

Table 4-1 shows the distribution of the risk outcome of any felony or violent misdemeanor arrest or revocation across the three time periods by supervision type and sex. In the first quarter of supervision (Period 1), men (13%) and women (12%) on straight probation had the highest outcome rate, compared to those on split probation

(9% for men, 10% for women) or parole (9% for men, 6% for women). In the next three quarters of the first year (Period 2), men (24%) and women (18%) on split probation terms were more likely to be arrested/revoked, compared to those on straight probation (21% for men, 18% for women) or parole (22% for men, 16% for women). For those that remained on supervision past one year (Period 3), individuals on split probation again were much more likely to experience the outcome (32% for men, 24% for women), compared to those on straight probation (24% for men, 19% for women) or parole (23% for men, 13% for women).

Table 4-1. Observed Rate of Outcome (Serious Arrest or Revocation) by Time Period, Supervision Type, and Sex

Variable	Period 1*			Period 2*			Period 3*		
	Straight Probation	Split Probation	Parole	Straight Probation	Split Probation	Parole	Straight Probation	Split Probation	Parole
Men N	62,218	27,169	14,079	50,240	22,655	11,407	29,723	13,125	5,419
Outcome**	12.7%	9.1%	8.7%	20.5%	23.8%	21.6%	24.1%	31.8%	22.9%
Women N	25,508	4,311	2,012	20,846	3,575	1,653	12,826	2,203	772
Outcome**	12.0%	9.6%	6.0%	18.3%	20.3%	15.7%	18.5%	24.4%	13.2%

* Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.
**Outcome is serious arrest or revocation and is reset at the beginning of each period.

Table 4-2 shows the distribution of outcome type (serious arrest or revocation) for those who experienced the outcome by period, supervision type, and sex. Serious arrest is broken down by charge type (drug, probation/parole, property, public order, violent) and charge severity (felony or misdemeanor). Although variations occur by supervision type, one notable similarity is that felony probation and parole charges often represent most of the outcomes in Periods 2 and 3. For men on supervision in the first quarter, in addition to probation/parole offenses, both drug and property charges are quite common. In Period 1, drug charges represent 22% of the outcomes for men on straight

probation, compared to 13% for men on straight probation and 16% for men on parole. Violent offenses (felony and misdemeanor combined) are also similarly prevalent in the first period, with little variation between men on straight and split probation (20%) but fewer violent arrests for men on parole (15%). For the remainder of the first year on supervision, felony probation/parole charges are the most common charge, representing 51% of serious arrests or revocations for men on straight probation, compared to 48% and 42% for those on split probation and parole, respectively. Similarly, drug and violent (both felony and misdemeanor violent) charges are the next most common charges in this period. In the remaining three quarters of the first year, drug charges make up 15% of outcomes for those on straight probation, compared to 14% for men on split probation and 20% for men on parole. Similarly, in this period, violent offenses represent 15% of outcomes for men on straight probation, in contrast with 18–19% for men on split probation or parole, respectively. This pattern persists for men after one year on supervision without a serious arrest or revocation. Notably, although rare, probation or parole revocation before a felony or violent arrest occurs in about 3% of outcomes in Period 2, and 3–4% in Period 3, for those on straight or split probation, compared to 8% for men on parole after one year on supervision.

Table 4-2. Outcome Type by Time Period, Supervision Type, and Sex

Outcome (Serious Arrest or Revocation) Type	Period 1*			Period 2*			Period 3*		
	Straight	Split	Parole	Straight	Split	Parole	Straight	Split	Parole
Men									
Felony drug	22%	13%	16%	15%	14%	20%	13%	13%	19%
Felony probation/parole	25%	35%	40%	51%	48%	42%	53%	51%	38%
Felony property	20%	17%	16%	10%	11%	11%	9%	8%	10%

Outcome (Serious Arrest or Revocation) Type	Period 1*			Period 2*			Period 3*		
	Straight	Split	Parole	Straight	Split	Parole	Straight	Split	Parole
Felony public order	13%	14%	11%	6%	6%	5%	6%	6%	5%
Felony violent	12%	11%	8%	7%	10%	10%	7%	9%	10%
Misdemeanor violent	8%	9%	7%	8%	8%	9%	10%	10%	11%
Revocation	1%	2%	2%	3%	3%	4%	3%	4%	8%
N	**62,218**	**27,169**	**14,079**	**50,240**	**22,655**	**11,407**	**29,723**	**13,125**	**5,419**
Men with outcome	**7,873**	**2,481**	**1,222**	**10,307**	**5,384**	**2,461**	**7,161**	**4,169**	**1,243**
				Women					
Felony drug	26%	14%	20%	16%	15%	20%	14%	12%	22%
Felony probation/parole	32%	41%	41%	61%	58%	52%	61%	58%	44%
Felony property	18%	14%	17%	9%	9%	11%	8%	9%	18%
Felony public order	13%	16%	10%	5%	4%	3%	4%	4%	0%
Felony violent	6%	5%	4%	2%	4%	4%	4%	6%	5%
Misdemeanor violent	5%	6%	6%	6%	7%	6%	7%	9%	8%
Revocation	1%	3%	2%	2%	3%	4%	3%	3%	3%
N	**25,508**	**4,311**	**2,012**	**20,846**	**3,575**	**1,653**	**12,826**	**2,203**	**772**
Women with outcome	**3,050**	**414**	**120**	**3,818**	**725**	**260**	**2,370**	**538**	**102**

* Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.

For women on supervision, these patterns are similar but have key differences. As shown in Table 4-2, in Period 1, a felony probation or parole charge is still the modal offense type, regardless of supervision type. However, violent charges are much less common for women than men. Instead, drug charges represent 26% of outcomes for women on straight probation, 20% for women on parole, and 14% for women on split probation. Felony property and public order offenses are similarly represented for women in Period 1. For property offenses, there is slight variation for women on straight (18%) and split probation (14%) compared to parole (17%). During Period 2, felony probation/parole charges represent most outcome types regardless of supervision type, with the next most common charge being drug charges, which comprise 15–20% of

outcome types. This pattern persists beyond one year of supervision, with felony probation/parole charges representing most outcomes for those on straight (61%) or split probation (58), and the modal charge for women on parole (44%).

As part of our analysis of arrest by time period, we also conducted survival analyses (described in Chapter 5).

Figure 4-1. Survival Hazard Function by Supervision Type

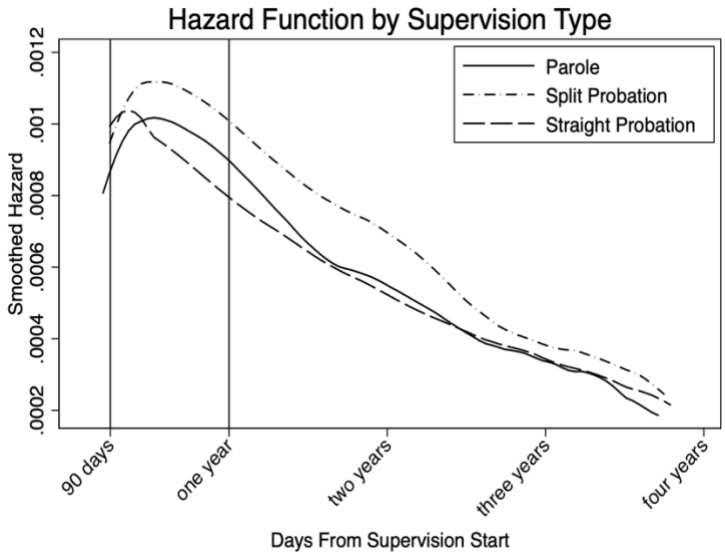


Figure 4-1 shows the smoothed hazard function of serious arrest or revocation by supervision type. The hazard function here shows that the risk of arrest rises from supervision start through a peak around the middle of the first year of supervision, before descending and dropping below the initial starting point after one year. The variation in serious arrest type and the differences in prevalence of risk of arrest in the different time periods, in addition to model test fit statistics, were part of the reasoning for using multiple and distinct time periods for modeling arrest while on supervision.

4.2 Static Factors

The static descriptive statistics are shown by sex and supervision type in Table 4-2. These statistics reveal substantial differences among the groups at the start of

supervision. Both men and women starting straight probation terms are younger than their counterparts on split probation or parole terms, with an average age of 34 years for both men and women on straight probation, compared to 35 and 38 years for split probation and parole, respectively. These demographic differences by supervision type persist, as 61% of men on straight probation are identified as "non-White" compared to slightly more than one-half of men on split probation and parole. Most women are also non-White, as 70% on straight probation, 66% on split probation, and 78% on parole are non-White. In addition, men on parole (20%) and on split probation (16%) are more likely to be identified as a confirmed gang member compared to those on probation (3%).[5] This pattern holds for women as well, but at much lower levels, with 4% of women on parole flagged as in a gang compared to 2% for those on split probation and 0.3% for those on straight probation.

Table 4-3. Static Descriptive Statistics by Supervision Type and Sex

Variable	Men			Women		
	Straight Probation n=62,218	Split Probation n=27,171	Parole n=14,079	Straight Probation n=25,508	Split Probation n=4,311	Parole n=2,012
Age (mean years)	33.63	35.15	38.10	33.97	35.32	37.77
Race = White	39.3%	49.7%	46.5%	29.6%	34.2%	21.7%
Race = non-White	60.7%	50.3%	53.5%	70.4%	65.8%	78.3%
Confirmed gang	2.9%	16.4%	21.3%	0.3%	1.9%	3.5%
Most Serious Current Charge*						
Violent	15.7%	40.8%	-	11.2%	26.1%	-
Property	25.7%	28.4%	-	32.4%	35.0%	-
Drug	41.2%	23.3%	-	43.0%	32.0%	-

[5] The confirmed gang measure is derived from both the GDC and the Georgia DCS. However, identification of gang membership is much more prevalent in the parole population because it includes prison gang membership. DCS also records information on gang membership, but this information is less easily confirmed once individuals are in the community; thus, it likely includes measurement error for those on straight probation.

Variable	Men			Women		
	Straight Probation n=62,218	Split Probation n=27,171	Parole n=14,079	Straight Probation n=25,508	Split Probation n=4,311	Parole n=2,012
Public order	8.2%	6.9%	-	6.3%	5.9%	-
Other	1.0%	0.5%	-	1.5%	1.0%	-
Unknown	8.2%	0.0%	-	5.6%	0.0%	-
Most serious offense for prison term was a property offense**	-	-	23.9%	-	-	30.3%
	-	-	76.1%	-	-	69.7%
Prison admission was for a revocation**	-	-	27.8%	-	-	29.7%
	-	-	72.2%	-	-	70.3%
Prior Arrests (any)						
Drug (past 2 years)	48%	27%	14%	48%	37%	24%
Drug (past 5 years)	56%	41%	43%	54%	48%	54%
Property (past 2 years)	38%	29%	10%	43%	39%	15%
Property (past 5 years)	48%	47%	36%	54%	57%	46%
Violent (past 2 years)	24%	24%	05%	17%	22%	05%
Violent (past 5 years)	33%	43%	26%	24%	35%	22%
Public order (past 2 years)	50%	42%	17%	42%	47%	22%
Public order (past 5 years)	62%	62%	52%	52%	63%	53%
Probation/parole (past 2 years)	15%	24%	23%	12%	23%	40%
Probation/parole (past 5 years)	25%	42%	49%	0.19	35%	59%
Any supervision (past 5 years)	12.6%	39.6%	48.6%	6.2%	27.7%	46.0%
Prison episode (past 2 years)*	4.1%	53.6%	-	1.4%	32.5%	-
Prison episode (past 5 years)*	6.6%	58.7%	-	2.3%	36.0%	-
In-prison disciplinary reports (mean)**	-	-	2.84	-	-	1.22
In-prison mental health treatment**	-	-	25.3	-	-	75.1

* Probation and split probation only.
**Parole only.

We identified the most serious current charge among all charges for those on probation, using the NCRP classifications and ordering violent, property, drug, and public order as the order of seriousness. The most common current charge for those on straight

probation was a drug offense (41% for men, 43% for women). For those on split probation, a violent charge (41%) was the most common charge for men and a property charge (35%) was the most common charge for women. For those on parole, 24% of men had been in prison for a property offense compared to 30% for women.[6] In this sample, less than one-third of men (28%) and women (30%) on parole had been in prison for a revocation.

The criminal history measures include counts of any arrest charges for specific offenses during limited lookback periods of 0–2 years and 0–5 years. These periods were determined to provide the best predictive value compared to longer lookback periods, including lifetime (see Chapter 6). Table 4-3 shows that most men (56%) and women (54%) on straight probation had one or more drug arrests, and roughly one-half of men (48%) and women (54%) had at least one property charge in the past 5 years. Violent charges were less prevalent among those on probation, with 33% of men and 24% of women having one or more violent charges in the 5 years before beginning probation. Public order charges were more common, experienced within the past 5 years by 62% of men and 52% of women on probation. Probation arrest charges were less common among those currently on probation; 25% of men and 19% of women had one or more probation/parole arrest charges in the previous 5 years.

For those on split probation, less than one-half had experienced a drug charge in the previous 5 years (41% of men and 48% of women) or a violent charge (43% of men and 35% of women). Property charges were somewhat more common (47% of men and

[6] Ongoing access to GDC data was limited to a selection of measures that are provided regularly to DCS that could be used to inform a dynamic risk assessment. One of these measures, which was incorporated in previous parole models for DCS, is whether the prison incarceration was for a property offense. That measure is retained here.

57% of women had had at least one in the previous 5 years). Public order charges were common, with 62% of the men and 63% of the women having one or more public order arrests in the previous 5 years. Probation/parole arrest charges were more common for those on split probation compared to straight probation, as 42% of men and 35% of women on split probation had one or more probation/parole arrests in the previous 5 years. Less than one-half of all groups had had a prior supervision in the previous 5 years.

For those on straight probation, most men (93%) and women (98%) had not had a prison episode in the 5 years before the start of probation. For those on split probation, most men (59%) and about one-third of women (36%) had had a prison episode in the previous 5 years. The final measures in Table 4-2 were based on GDC data and reflected measures captured during the prison term preceding the parole. Men had an average of 2.84 disciplinary reports, and women had an average of 1.22 disciplinary reports, during the prison term before their parole. About one-fourth (25%) of men and three-fourths (75%) of women had a record of mental health issues while in prison.

Continuing with a description of the recent criminal history of the cohort, Table 4-4 shows the mean number of arrests before the start of supervision by arrest type for individuals who had at least one arrest of that type. Variation by supervision type is usually negligible, depending on the charge type. For example, the mean number of drug arrests for those who had at least one ranges between 1.4 and 1.6 for all supervision types and for both men and women, despite much of the cohort having at least one drug arrest in the past 5 years. Public order arrests have some of the highest mean counts for those who have at least one public order charge, with people reporting

roughly 2 public order arrests across supervision type and sex. However, some variation exists for those who have had at least one probation/parole charge in the past 5 years. While men with these charges all have about 2 charges across supervision types, women on parole have the highest mean counts, with 2.5 probation/parole charges, compared to 1.8 for women on straight probation.

Table 4-4. Static Criminal History Continuous Measures by Supervision Type and Sex

Variable	Men			Women		
	Straight n=62,218	Split n=27,171	Parole n=14,079	Straight n=25,508	Split n=4,311	Parole n=2,012
Mean Number of Arrests Among Individuals with at Least 1 Prior Arrest						
Drug arrests past 2 years	1.3	1.3	1.1	1.3	1.3	1.2
Probation/parole arrests past 2 years	1.5	1.4	1.3	1.4	1.5	1.4
Property arrests past 2 years	1.4	1.5	1.2	1.4	1.6	1.3
Public order arrests past 2 years	1.6	1.6	1.3	1.5	1.6	1.4
Violent arrests past 2 years	1.2	1.2	1.1	1.2	1.2	1.1
Drug arrests past 5 years	1.5	1.6	1.5	1.4	1.6	1.6
Probation/parole arrests past 5 years	2.0	2.1	2.2	1.8	2.1	2.5
Property arrests past 5 years	1.7	1.9	1.8	1.7	2.0	1.9
Public order arrests past 5 years	2.1	2.2	1.9	1.9	2.1	2.0
Violent arrests past 5 years	1.4	1.5	1.3	1.3	1.4	1.2

4.3 Dynamic Factors

Table 4-5 provides the results for men for the dynamic measures that were allowed to change over the course of supervision. Variations occurred across supervision types as well as across the three periods.

In the first period, the percentage of men on parole who had at least one positive drug test (15%) was higher compared to the percentage of men with at least one positive test

on straight (10%) or split probation (7%). (The models include counts of positive and negative tests from the beginning of supervision; here, we show the percentage of men with at least one.) This pattern held over the next two time periods, with all groups more likely to have a positive test over the longer periods at risk during the second two periods. Nearly one-half (48%) of men on parole had at least one negative drug test in the first 90 days compared to 23% and 20% on straight and split probation, respectively. This same ordering followed for Periods 2 and 3, but at higher rates, which was likely at least partially because of the longer time in the subsequent periods. Notably, individuals on probation or parole could have both positive and negative tests recorded during the period. For example, on Day 2, an individual could test for at least one substance and the resulting positive drug test measure would increment by one. Similarly, if the same individual was tested on Day 3 and all substance results were null, the negative drug test count would increment by one, while the positive drug test value would remain the same as it had on Day 2. Men on parole also had higher prevalence rates compared to those on straight and split probation for technical violations.

Non-violent misdemeanor arrests were included in the models as covariates. Table 4-5 shows that the men had few non-violent misdemeanor arrests of different types for all supervision types. Public order misdemeanor arrests were the most frequent, with 3% of men having a public order arrest in Period 1, regardless of the supervision type, roughly 7% of men in Period 2, and approximately 11% in Period 3. Probation special conditions observed in Periods 1 and 2 had a much higher prevalence rate across categories. Drug and alcohol restrictions were the most common in the first 90 days, as 54% of men on straight probation had a special condition compared to 46% of males on split

probation. This condition was also the most common in Period 2. Drug or alcohol treatment conditions were the second-most common condition, as nearly one-third of men on probation had a condition of that type in Period 1. Fee-related conditions were also quite prevalent, as more than 40% of men on straight or split probation had a special condition for fees in the first 90 days.

Table 4-5. Dynamic Measures for Men by Supervision Type

Variable	Period 1			Period 2			Period 3		
	Straight n=62,218	Split n=27,171	Parole n=14,079	Straight n=25,508	Split n=4,311	Parole n=2,012	Straight n=62,218	Split n=27,171	Parole n=14,079
Counter Variables (percentage with any)									
Any misdemeanor drug arrests	1%	1%	0%	2%	2%	1%	3%	3%	2%
Any misdemeanor parole/probation arrests	2%	1%	0%	3%	2%	0%	4%	3%	1%
Any misdemeanor property arrests	1%	1%	1%	3%	3%	2%	3%	3%	2%
Any misdemeanor public order arrests	3%	3%	3%	8%	7%	6%	11%	11%	10%
Any negative drug tests	23%	20%	48%	44%	38%	64%	60%	52%	72%
Any positive drug tests	10%	7%	15%	19%	15%	24%	25%	21%	28%
Counter Variables (mean of the number of observed events)									
Count of changes of supervision level*	1.9	1.7	1.4	1.4	1.4	1.3	1.5	1.5	1.3
Count of the number of violations	0.2	0.1	0.4	0.3	0.2	0.8	0.4	0.3	0.9
Dichotomous Variables (percentage with this value on the last observed day in the period)									
No drug tests in last 90 days	-	-	-	81%	84%	75%	88%	89%	82%
Violence-related probation condition	19%	16%	-	16%	14%	-	-	-	-
Community service probation condition	26%	18%	-	22%	15%	-	-	-	-
Drug- or alcohol-related probation condition	54%	46%	-	46%	40%	-	-	-	-
Drug or alcohol treatment probation condition	33%	31%	-	28%	26%	-	-	-	-

Variable	Period 1			Period 2			Period 3		
	Straight n=62,218	Split n=27,171	Parole n=14,079	Straight n=25,508	Split n=4,311	Parole n=2,012	Straight n=62,218	Split n=27,171	Parole n=14,079
Education-related probation condition	7%	9%	-	6%	7%	-	-	-	-
Employment probation condition	3%	4%	-	3%	4%	-	-	-	-
Fee-related probation condition	41%	42%	-	36%	36%	-	-	-	-
No contact order probation condition	25%	34%	-	21%	29%	-	-	-	-
Other probation condition	42%	36%	-	35%	30%	-	-	-	-
Any parole condition	-	-	82%	-	-	63%	-	-	-
Warrant supervision level active	5%	5%	4%	17%	20%	14%	22%	28%	15%
Contact supervision level active	1%	1%	-	5%	3%	-	30%	22%	8%
Employment indicator	30%	25%	34%	35%	29%	37%	-	-	-

*Count resets at the beginning of each new model period.

Table 4-6 shows the average counts of dynamic counter variables for men who had at least one of the indicated events. Notably, for those with at least one misdemeanor arrest (regardless of arrest type), variation by supervision type or period is limited. For those who have at least one misdemeanor arrest while on supervision, the average number of misdemeanor arrests ranges from 1.0 to 1.3 across supervision types, time periods, and arrest types. However, variation by time period is substantial for drug testing. In the first period, for those who had at least one drug test of the type, the average number of negative drug tests was roughly 1.7, and the average number of positive tests was approximately 1.36, across supervision types. In the second period, the average negative drug test count ranged from 2.5 to 2.9, whereas the average positive tests ranged between 1.7 and 1.9 by type. After one year, for men who had a

negative drug test, parolees had on average 4.5 negative tests compared to 3.7 for men on split probation and 4.1 for men on straight probation.

Table 4-6. Dynamic Count Measures for Men by Supervision Type and Time Period

Variable	Period 1			Period 2			Period 3		
	Straight n=62,218	Split n=27,171	Parole n=14,079	Straight n=25,508	Split n=4,311	Parole n=2,012	Straight n=62,218	Split n=27,171	Parole n=14,079
Mean for Individuals with at Least 1 Event									
Misdemeanor drug arrests	1.0	1.0	1.0	1.1	1.1	1.1	1.1	1.1	1.1
Misdemeanor parole/ probation arrests	1.0	1.0	1.0	1.1	1.1	1.1	1.2	1.1	1.1
Misdemeanor property arrests	1.1	1.1	1.0	1.1	1.2	1.1	1.2	1.2	1.1
Misdemeanor public order arrests	1.1	1.0	1.0	1.1	1.1	1.1	1.2	1.2	1.3
Negative drug tests	1.7	1.6	1.7	2.9	2.5	2.9	4.1	3.7	4.5
Positive drug tests	1.4	1.3	1.4	1.8	1.7	1.9	2.1	2.0	2.4

Table 4-7 shows the values of the dynamic indicators for the women on probation and parole during the three periods. These patterns were often different when compared to those identified in the male population. For instance, women on parole were more likely to have a positive drug test (7%) than those on split probation (6%) in Period 1, but less likely compared to those on straight probation (11%). This pattern also persisted throughout the three time periods. Interestingly, women on parole were much less likely to have had at least one negative drug test throughout the three time periods (ranging from 46–75%) compared to split probation (21–56%) and straight probation (26–62%).

Yet, like men on parole, women on parole were more likely to have at least one technical violation during the time periods (ranging from 25–61%) compared to split (16–32%) and straight probation (20–38%). Like the men, few women had misdemeanor arrests of any type during supervision. However, public order misdemeanor arrests were the most common. In Period 1, 3% of women on both split and straight probation recorded a misdemeanor arrest, compared to women on parole (<1%). Subsequently, in Period 2, more women on straight (6%) and split probation (7%) had at least one public order misdemeanor arrest compared to women on parole (3%), with this pattern persisting past one year on supervision. After the first period, a substantial percentage of the female cohort did not register a drug test in the prior 90 days before the day of their last observation, with more than three-quarters of the cohort not having been recently tested, regardless of supervision type. This pattern persisted and increased after one year on supervision without a felony arrest, violent misdemeanor arrest, or revocation.

Table 4-7. Dynamic Measures for Women by Supervision Type

Variable	Period 1			Period 2			Period 3		
	Straight (19,864)	Split (3,392)	Parole (1,562)	Straight (16,209)	Split (2,799)	Parole (1,269)	Straight (9,972)	Split (1,718)	Parole (576)
Counter Variables (percentage with any)									
Any misdemeanor drug arrests	0%	0%	0%	1%	1%	1%	1%	2%	1%
Any misdemeanor parole/probation arrests	2%	2%	0%	3%	3%	1%	4%	3%	1%
Any misdemeanor property arrests	1%	2%	0%	3%	4%	2%	4%	4%	3%
Any misdemeanor public order arrests	3%	3%	1%	6%	7%	3%	8%	8%	6%
Any negative drug tests	26%	21%	46%	47%	39%	64%	62%	56%	75%

Variable	Period 1			Period 2			Period 3		
	Straight (19,864)	Split (3,392)	Parole (1,562)	Straight (16,209)	Split (2,799)	Parole (1,269)	Straight (9,972)	Split (1,718)	Parole (576)
Any positive drug tests	11%	6%	7%	19%	14%	17%	23%	20%	21%
Counter Variables (mean of the observed events)									
Number of changes of supervision level*	1.8	1.7	1.4	1.4	1.4	1.3	1.5	1.6	1.3
Number of violations	0.2	0.2	0.3	0.3	0.2	0.6	0.4	0.3	0.6
Dichotomous Variables (percentage with this value on the last observed day in the period)									
No drug tests in last 90 days	-	-	-	79%	83%	75%	89%	91%	85%
Violence-related probation condition	17%	14%	-	14%	12%	-	-	-	-
Community service probation condition	27%	25%	-	23%	21%	-	-	-	-
Drug- or alcohol-related probation condition	54%	53%	-	45%	45%	-	-	-	-
Drug or alcohol treatment probation condition	34%	38%	-	27%	32%	-	-	-	-
Education-related probation condition	7%	8%	-	6%	6%	-	-	-	-
Employment probation condition	2%	3%	-	2%	3%	-	-	-	-
Fee-related probation condition	42%	46%	-	36%	39%	-	-	-	-
No contact order probation condition	24%	33%	-	21%	29%	-	-	-	-
Other probation condition	44%	39%	-	36%	32%	-	-	-	-
Any parole condition	-	-	83%	-	-	58%	-	-	-
Warrant supervision level active	5%	5%	4%	17%	20%	12%	18%	25%	9%
Contact supervision level active	1%	1%	-	8%	5%	-	37%	32%	15%
Employment indicator	26%	22%	30%	30%	27%	32%	-	-	-

*Count of supervision-level changes resets at the beginning of each new model period.

Differences occur in the percentages with each of the probation conditions, although there is little variation between the women on straight or split probation. In addition, the prevalence of each condition type is often lower after the first period, as more individuals with the conditions are likely to have been arrested. In the first period,

roughly one in four women on probation have a community service condition, and more than one-half of women on probation have a drug or alcohol restriction condition. Yet only 34% and 38% of women on straight and split probation had a drug or alcohol treatment condition imposed at sentencing and present for the first period of supervision. Moreover, 42–46% of women on straight and split probation had a fee-related condition in Period 1. Lastly, the percentage of women who were reported to be employed on the last observed day in the first period was 26% for women on straight probation, compared to 22% and 30% for those on split probation and parole.

Table 4-8 shows the means for a selection of dynamic counter measures for women on supervision who had at least one of the events occur (e.g., the mean of misdemeanor drug arrests while on supervision for women who had at least one misdemeanor drug arrest). These factors follow a similar, if attenuated, pattern to that observed with the men. Misdemeanor arrests of any type are relatively infrequent, as the mean of any arrest type by time period and supervision type hovers around 1.0. Like the male cohort, variety is greater for positive and negative drugs throughout the three time periods. In Period 1, for women who have at least one negative drug test, the average is roughly 1.7 compared to 1.5 for positive tests for women with at least one positive test. In Period 2, the average counts of negative drug tests increase to 2.6, 3.0, and 3.1 for women on split probation, straight probation, or parole. The average number of positive drug tests for this period is 1.8 for women on straight or split probation and 2.1 for women on parole. After one year on supervision, the averages for those who have

recorded a negative drug test are higher still, as women on split probation have an average of 3.8 negative tests compared to 4.2 for straight probation and 4.3 for women on parole.

Table 4-8. Dynamic Count Measures for Women by Supervision Type and Time Period

Variable	Period 1			Period 2			Period 3		
	Straight (19,864)	Split (3,392)	Parole (1,562)	Straight (16,209)	Split (2,799)	Parole (1,269)	Straight (9,972)	Split (1,718)	Parole (576)
Mean Number for Individuals with at Least 1 Event									
Misdemeanor drug arrests	1.0	1.1	1.0	1.1	1.1	1.1	1.1	1.1	1.0
Misdemeanor parole/probation arrests	1.1	1.0	1.0	1.1	1.1	1.0	1.2	1.1	1.0
Misdemeanor property arrests	1.1	1.1	1.0	1.1	1.1	1.1	1.1	1.2	1.0
Misdemeanor public order arrests	1.1	1.0	1.0	1.1	1.1	1.0	1.2	1.2	1.1
Negative drug tests	1.7	1.7	1.8	3.0	2.6	3.1	4.2	3.8	4.3
Positive drug tests	1.4	1.5	1.6	1.8	1.8	2.1	2.1	2.2	2.2

4.4 Cohort Characteristics Summary

Overall, in the cohort of individuals starting felony probation or parole between 2016 and 2019, roughly 39% are arrested for a felony or violent misdemeanor arrest or have their supervision revoked. Differences by supervision type and sex are notable in the outcome of felony or violent misdemeanor arrest or revocation, yet there is more similarity across groups within modeling time periods. In the first quarter of supervision, arrest rates generally range from 6–13%. In contrast, given the increased exposure time, the arrest rates for the remaining three quarters of the first year range from 16–24%. Yet after one year on supervision without an arrest or revocation, variation is

substantial in rearrest rates. This variation persists when examining the outcome by arrest type. In the first period, non-probation-related charges (i.e., new criminal activity) are more common. However, after the first quarter on supervision, felony probation/parole arrests, which include arrests for technical violations or other finger-printable charges, represent the majority or modal arrest type across supervision type and sex.

Differences by supervision type and sex are substantive in both static and dynamic measures associated with arrest during supervision. Recent criminal history varies by group, but is also driven by exposure, as individuals on parole or split probation often had less-recent criminal history compared to those on straight probation, because of the nature of their prison sentences. Once on supervision, key substantive differences are also shown in dynamic measures by groups. While misdemeanor arrests are rare across the groups, the results of drug tests vary substantially by supervision type and sex, as both men and women on parole are much more likely to test negative for drugs compared to those on probation. Special probation conditions are also common across sex and supervision types, despite variations in condition application; nearly one-half of the probation cohort includes a condition related to drugs or alcohol, whereas education and employment conditions are present in less than 10% of each group. This descriptive variation by group, sex, and period provides further support for producing separate models by these characteristics, which is discussed in the next chapter.

5. Final Models of Risk of Arrest or Revocation

This chapter presents the final models that were produced through the development process described in Chapter 6. The set of risk assessment models improves upon the predictive accuracy of the existing set of algorithms in use by DCS. As noted previously, DCS requested that serious arrest or revocation be used as the outcome, where serious arrest is defined as a felony or violent misdemeanor arrest. Following the current practice at DCS and reflecting differences in risk profiles, separate models were developed for men and women and for three supervision types (straight probation, split probation, and parole). In addition, examining changes in risk profiles over supervision periods resulted in an assessment that the best fit was provided by final sets of models for three consecutive time periods: the first quarter of supervision (Period 1), the next three quarters the first year of supervision (Period 2), and one-plus years of supervision (Period 3).

5.1 Final Specifications for Models for Men

Table 5-1 shows the odds ratios from the final set of logistic regression models for the static and dynamic factors for men on straight or split probation or parole during Periods 1 through 3. As per usual interpretation of odds ratios, values less than 1 denote reduced odds of the outcome, while those greater than 1 indicate increased odds. Odds ratios significantly different from 1 at the $p < 0.001$ level are shown in bold. The 0.001 significance level was chosen because of the very large sample sizes for most models.

Table 5-1. Model Results for Men by Supervision Type and Period

Factor	Odds Ratios								
	Period 1*			Period 2*			Period 3*		
	Straight (48,513)	Split (21,149)	Parole (10,988)	Straight (39,220)	Split (17,645)	Parole (8,893)	Straight (23,225)	Split (10,216)	Parole (4,236)
(Intercept)	**0.29**	**0.23**	**0.08**	**0.25**	**0.39**	**0.13**	**0.39**	**0.54**	0.82
Age at start of supervision	**0.99**	**0.99**	**0.98**	**0.99**	**0.99**	**0.98**	**0.99**	**0.99**	**0.98**
Confirmed gang	**1.89**	**1.36**	**1.81**	**2.35**	**1.48**	**1.72**	**2.95**	**1.42**	**2.13**
Prior Arrests**									
Violent offenses	1.05	1.06	1.06	**1.21**	1.08	**1.24**	**1.32**	**1.23**	0.93
Public order offenses	**1.06**	1.01	1.08	**1.05**	1.02	1.06	**1.07**	1.02	0.88
Drug offenses	**1.08**	1.07	0.95	**1.15**	1.03	1.09	**1.22**	1.04	1.32
Property offenses	**1.12**	**1.10**	**1.15**	**1.17**	**1.07**	**1.16**	**1.22**	1.05	0.84
Probation/parole offenses	**1.10**	**1.12**	**1.10**	**1.13**	**1.14**	**1.13**	**1.22**	**1.25**	0.78
Prior prison terms**	**0.76**	**0.75**	--	**1.26**	**1.20**	--	**1.69**	**1.24**	--
Prior probation or parole (past 5 years)	**1.29**	1.03	**1.51**	1.08	1.07	**1.27**	**1.24**	**1.29**	1.22
Drug Testing***									
Positive drug tests	**1.30**	**1.35**	**1.23**	**1.13**	**1.19**	**1.11**	**1.05**	**1.10**	1.06
Negative drug tests	**0.76**	0.89	**0.63**	**0.95**	**0.90**	**0.91**	**0.98**	0.99	**0.97**
No drug testing in previous 90 days	--	--	--	**0.63**	**0.52**	**0.37**	**0.37**	**0.34**	**0.32**
Employed	**0.51**	**0.58**	**0.47**	**0.76**	0.87	**0.70**	--	--	--
Violations during supervision (any type)	1.03	1.05	**1.18**	1.03	1.04	**1.13**	**1.09**	1.06	**1.12**
Active warrant	**8.47**	**6.71**	**4.31**	**7.47**	**6.12**	**6.67**	**6.90**	**5.99**	**10.06**
Moved to contact status	0.43	0.18	--	**0.28**	**0.46**	--	**0.33**	**0.35**	**0.30**
Count of supervision changes during period	**0.45**	**0.54**	**0.61**	**0.52**	**0.55**	**0.50**	**1.15**	**1.19**	1.02
Most Serious Charge on Docket (ref: drug)									
Other offenses	0.74	0.63	--	0.84	0.73	--	0.54	0.66	--
Property offenses	1.04	1.03	--	1.00	1.10	--	0.91	1.01	--
Public order offenses	0.87	1.14	--	1.04	1.06	--	0.93	0.84	--
Violent offenses	**0.80**	**0.74**	--	**0.83**	0.97	--	1.00	0.84	--
Missing	**6.07**	--	--	**2.08**		--	**1.51**		--
Probation Conditions									
Community service	1.11	1.12	--	**1.16**	**1.34**	--	--	--	--

Factor	Odds Ratios								
	Period 1*			Period 2*			Period 3*		
	Straight (48,513)	Split (21,149)	Parole (10,988)	Straight (39,220)	Split (17,645)	Parole (8,893)	Straight (23,225)	Split (10,216)	Parole (4,236)
Drug or alcohol restrictions	1.10	1.12	--	**1.52**	**1.25**	--	--	--	--
Drug or alcohol testing	**1.27**	1.21	--	**1.39**	**1.41**	--	--	--	--
Education	1.11	1.08	--	**1.46**	1.03	--	--	--	--
Employment	1.00	1.08	--	1.11	1.36	--	--	--	--
Fees	**1.19**	1.19	--	**1.37**	**1.37**	--	--	--	--
No contact orders	**1.18**	**1.37**	--	**1.30**	**1.32**	--	--	--	--
Other	**1.28**	1.16	--	**1.37**	**1.52**	--	--	--	--
Violence-related conditions	1.10	1.03	--	**1.14**	1.16	--	--	--	--
Misdemeanor Arrests During Supervision									
Drug offenses	1.23	1.41	0.98	**1.59**	**1.89**	0.88	**1.67**	**1.95**	1.06
Public order offenses	1.19	1.08	1.14	**1.50**	**1.63**	1.32	**1.48**	**1.69**	**2.14**
Property offenses	**1.46**	**2.46**	**2.28**	**2.22**	**2.15**	**2.12**	**1.93**	**1.73**	1.40
Probation/parole offenses	**0.50**	1.03	0.92	1.19	1.21	1.43	1.22	1.13	0.52
Parole-Specific Measures from GDC									
Primary offense is for property charge	--	--	1.16	--	--	**1.38**	--	--	1.11
Prison admission was for revocation	--	--	1.01	--	--	1.19	--	--	1.09
Count of prison disciplinary reports	--	--	1.01	--	--	1.01	--	--	1.00
Mental health treatment flag	--	--	**1.50**	--	--	**1.37**	--	--	**1.47**
Any parole condition	--	--	**3.06**	--	--	**13.03**	--	--	--

* Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.
**For prior arrests and prior prison/supervision, the Period 1 and 2 models use 5 years as the lookback period, and the Period 3 models use 2 years as the lookback period.
***Any positive test on a day is recorded as a positive test; a negative test on a day is recorded if all tests are negative.

Nearly all factors consistently predict either increased or reduced risk across the models. For example, factors associated with increased risk are consistently associated with increased risk across the models. Consistent with most recidivism findings, age is associated with reduced risk in each model, in this model operationalized as a

continuous measure. Confirmed gang membership is a risk factor associated with a higher likelihood of arrest/revocation for all models.

The static criminal history measures are counts of prior arrest charges by offense type during limited lookback periods (previous 5 years for the Period 1 and 2 models; previous 2 years for the Period 3 models, based on testing of predictive value for various lookback periods, as described later in this chapter). Prior arrests are positive predictors of the outcome, although the effects are small to negligible for some factors in some of the models. A record of a prison term in the 5 years before the start of the probation term is a risk factor during the initial 90-day period for both straight and split probation but is protective thereafter. (Prior prison is not included in the parole models because all on parole will have had a recent prison term.) A record of a previous probation or parole term in the previous 5 years is generally associated with increased risk of the outcome, although not always significant at the 0.001 level.

Three drug test measures are included in the models. Counts of positive and negative tests during each period are included in all models. In addition, at the recommendation of DCS, an indicator of whether an individual was tested in the previous 90 days was included in the Period 2 and 3 models. Positive drug tests are associated with increased risk of rearrest/revocation throughout, although this effect is small in the Period 3 models. For example, for men on straight probation, one positive drug test increases the odds of rearrest/revocation by 30% in the first 90 days of supervision, whereas this same test result only increases the odds of rearrest by 5% after one year on supervision without a rearrest. This finding also holds for men on split probation and parole, with a positive drug test resulting in increased odds of rearrest/revocation. Negative drug tests

(i.e., test results for all substances are negative in a day) are associated with a reduced risk of arrest/revocation, with the effect the strongest in the first 90 days, when more and more random testing is conducted during the initial period of supervision (i.e., first 90 days). After 90 days, drug testing is not random and is commonly administered based on suspicion of drug use, leading to fewer overall tests and fewer negative tests (potentially explaining the lack of significance past one year). Finally, the “no testing in the previous 90 days” factor is highly significant and indicates reduced risk in the latter two periods for all three supervision types.

Employment was associated with a reduced risk of rearrest/revocation during the first year of supervision but could not be included in the >1-year model because the data were not reliably verified and entered after the initial year. Not surprisingly, violations were associated with increased risk, and having an active warrant was a strong indicator of risk across the models. Contact status is a reduction in supervision intensity that incorporates telephone, as opposed to in-person, contact and a reduced frequency of check-ins. For those on probation, having the supervision level reduced from a standard to contact status as a result of compliance with conditions is associated with reduced risk of arrest/revocation. The number of supervision changes during the period was protective during the first year, but was largely an indicator of increased risk after the first year—likely because changes in the initial year were reductions in supervision level in response to compliance to contact, and those later were increases in intensity because of issues.

For those on probation, the most serious charge associated with the term is generally not predictive of risk. Compared to having a drug charge as the most serious charge,

having a property, public order, or other offense as the most serious charge reflected similar risk. However, those on probation whose most serious charge was for a violent offense were less likely than those with a drug charge as the most serious charge to have an arrest/revocation during the first year. The most serious charge was not available for those on parole, but two factors were included in the parole models that suggested slightly higher risk for those on parole whose prison term had been for a property offense or a probation/parole violation.

Probation conditions were generally associated with increased risk of arrest/revocation during the first year of probation. Conditions that extended beyond the first year were almost universally associated with increased risk of the outcome and could not be included in the Period 3 models.

Misdemeanor arrests for non-violent charges were not included in the outcome, and thus could be included as risk factors in the models. Counts of misdemeanor arrests were included in all models. Not surprisingly, property, drug, and public order arrests signaled an increased risk of future arrest/revocation.

For individuals on parole supervision, information from their prison term included an indicator of mental health treatment. This variable was included in the three parole models and signaled an increased risk of arrest/revocation. Having any condition attached to parole was also associated with an increased risk of arrest/revocation.

5.2 Final Specifications for Models for Women

Table 5-2 shows the odds ratios from the final logistic regression models for the static and dynamic factors for the risk algorithms for women on either straight or split probation or parole for the three periods.

Table 5-2. Model Results for Women by Supervision Type and Period

Factor	Odds Ratios*								
	Period 1			Period 2			Period 3		
	Straight (19,864)	Split (3,392)	Parole (1,562)	Straight (16,209)	Split (2,799)	Parole (1,269)	Straight (9,972)	Split (1,718)	Parole (576)
(Intercept)	**0.403**	**0.261**	0.078	**0.373**	**0.331**	**0.049**	**0.309**	0.639	0.166
Age at start of supervision	**0.988**	**0.976**	0.963	**0.987**	**0.975**	0.998	**0.990**	0.980	0.995
Confirmed gang	1.194	1.246	0.810	2.833	1.829	2.444	1.502	2.052	3.059
Prior Arrests									
Violent offenses	1.048	1.210	1.006	**1.244**	0.998	0.890	**1.407**	1.359	0.967
Public order offenses	**1.101**	1.040	1.153	**1.082**	1.029	1.060	1.058	1.030	0.907
Drug offenses	1.009	1.196	0.987	**1.122**	**1.309**	1.054	**1.294**	1.101	1.061
Property offenses	**1.122**	1.133	1.043	**1.167**	**1.174**	1.184	1.122	1.185	1.249
Probation/parole offenses	**1.128**	1.108	0.998	1.077	**1.176**	**1.344**	**1.308**	**1.423**	0.625
Prior prison terms (0–5 years)	0.672	**0.443**	--	0.634	1.164	--	1.766	1.533	--
Prior probation or parole (0–5 years)	1.141	1.202	1.923	1.222	1.085	1.120	1.025	1.502	1.367
Drug Testing									
Positive drug tests (any positive test on date)	**1.476**	**1.628**	**1.695**	**1.210**	**1.309**	**1.418**	**1.072**	1.100	1.352
Negative drug tests (all tests negative on date)	**0.767**	0.768	0.555	**0.938**	0.928	**0.825**	**0.977**	0.994	0.985
No drug testing in previous 90 days	--	--	--	**0.521**	0.666	**0.278**	**0.305**	**0.296**	0.291
Employed	**0.430**	**0.433**	**0.319**	**0.631**	**0.594**	0.616	--	--	--
Violations during supervision (any type)	0.988	0.904	1.035	0.985	1.008	1.039	1.028	1.066	1.042
Active warrant	**8.262**	**8.378**	**7.580**	**10.472**	**8.434**	**8.157**	**10.058**	**7.189**	**40.459**
Moved to contact status	0.691	**1.492**	--	**0.253**	0.331	--	**0.324**	**0.436**	1.180
Count of supervision-level changes during period	**0.431**	0.549	**0.410**	**0.484**	**0.468**	0.563	**1.217**	1.054	0.834
Most Serious Charge on Docket (ref: drug)									
Other offenses	1.183	0.906	--	1.020	0.771	--	0.483	1.472	--
Property offenses	0.844	1.190	--	**0.774**	1.093	--	0.939	0.850	--
Public order offenses	0.741	0.934	--	0.717	1.265	--	0.660	0.797	--
Violent offenses	0.745	0.763	--	0.738	1.150	--	0.938	0.861	--
Missing	**6.184**	--	--	**1.942**	--	--	**2.033**	--	--

Factor	Odds Ratios*								
	Period 1			Period 2			Period 3		
	Straight (19,864)	Split (3,392)	Parole (1,562)	Straight (16,209)	Split (2,799)	Parole (1,269)	Straight (9,972)	Split (1,718)	Parole (576)
Probation Conditions									
Community service	1.068	1.140	--	1.129	**1.894**	--	--	--	--
Drug or alcohol restrictions	1.157	1.150	--	**1.439**	1.200	--	--	--	--
Drug or alcohol testing	**1.406**	1.319	--	**1.627**	**1.578**	--	--	--	--
Education	1.035	1.418	--	1.169	1.247	--	--	--	--
Employment	1.064	1.011	--	0.985	1.123	--	--	--	--
Fees	1.176	1.302	--	**1.286**	1.357	--	--	--	--
No contact orders	**1.241**	1.376	--	**1.465**	1.303	--	--	--	--
Other	**1.287**	1.496	--	**1.338**	1.281	--	--	--	--
Violence-related conditions	1.202	0.840	--	1.055	1.043	--	--	--	--
Misdemeanor Arrests During Supervision									
Drug offenses	2.267	2.326	--	1.622	1.513	3.024	0.983	1.315	2.898
Public order offenses	0.855	0.758	2.076	**1.553**	1.655	0.734	**1.812**	1.819	5.313
Property offenses	1.504	1.417	16.489	**2.269**	**2.544**	2.068	**2.110**	1.764	1.927
Probation/parole offenses	**0.418**	0.778	0.907	1.300	1.416	0.650	1.322	1.297	0.759
Parole-Specific Measures from GDC									
Primary offense is for property charge	--	--	0.892	--	--	0.629	--	--	0.449
Prison admission was for revocation	--	--	1.333	--	--	0.909	--	--	2.765
Count of prison disciplinary reports	--	--	1.050	--	--	1.032	--	--	1.001
Mental health treatment flag	--	--	1.485	--	--	1.133	--	--	1.397
Any parole condition	--	--	**6.864**	--	--	**16.840**	--	--	--

* Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.
**For prior arrests and prior prison/supervision, the Period 1 and 2 models use 5 years as the lookback period, and the Period 3 models use 2 years as the lookback period.
***Any positive test on a day is recorded as a positive test; a negative test on a day is recorded if all tests are negative.

As with the men's models, nearly all factors are consistently either protective or risk factors across the models. The direction of the relationships is also generally consistent

with the results of the men's models, although some differences occur. As sample sizes are smaller for the female subpopulations, fewer odds ratios are statistically different from 1.0 at the 0.001 level. Consistent with most recidivism findings, age is a protective factor in each model. Confirmed gang membership is a risk factor associated with a higher likelihood of arrest/revocation for all models.

Counts of prior arrest charges by offense type during limited lookback periods (previous 5 years for the Period 1 and 2 models; previous 2 years for the Period 3 models) are generally associated with higher risks of the outcome, although the effects are small to negligible for some factors in some of the models, and prior violent charges are associated with a reduced risk during Period 2 for women on split probation or parole. In contrast to the findings for the men's models, a record of a prison term in the 5 years before the start of the probation term is associated with reduced risk during the initial 90-day period for those on straight and split probation, and during Period 2 for those on straight probation, but is a risk factor otherwise. (Prior prison is not included in the parole models because all on parole will have had a recent prison term.) A record of a probation or parole term in the previous 5 years is associated with increased risk of the outcome.

Results for the three drug test measures mirror those for the men's models. Positive drug tests are associated with increased risk of rearrest/revocation throughout, and negative drug test results are associated with reduced risk. The "no testing in the previous 90 days" factor is highly significant and indicates reduced risk in the latter two periods for all three supervision types.

Employment was associated with a reduced risk of rearrest/revocation during the first year of supervision. In contrast to the men's models, violations were only slightly associated with increased risk, although having an active warrant was a strong indicator of risk across the models. Results were like those for the men with respect to changes in supervision level—movement to contact supervision level was associated with reduced risk throughout, while the number of supervision changes during the period was protective during the first year but largely an indicator of increased risk after the first year. This difference was likely because changes in the initial year were reductions in supervision level in response to compliance, and changes later were increases in intensity because of issues.

For those on probation, compared to having a drug charge as the most serious charge, those with a property, public order, or other offense as the most serious charge generally had lower—although not significantly lower—risk. The most serious charge was not available for those on parole, but two factors were included in the parole models that suggested slightly lower risk for those on parole whose prison term had been for a property offense (the opposite effect compared to the men's models), and a higher risk if the term was for a probation/parole violation.

Like the men's models, probation conditions were associated with increased risk of arrest/revocation during the first year of probation. Conditions that extended beyond the first year were almost universally associated with the outcome and could not be included in the >1-year models.

Misdemeanor arrests for non-violent charges were not included in the outcome, and thus could be included as risk factors in the models. Counts of misdemeanor arrests

were included in all models. Not surprisingly, property, drug, and public order arrests signaled an increased risk of future arrest/revocation—like in the men's models.

For individuals on parole supervision, information from their prison term included an indicator of mental health treatment. This variable was included in the three parole models and signaled an increased risk of arrest/revocation. Having any condition attached to parole was also associated with an increased risk of arrest/revocation.

6. Model Development

This chapter describes the model development and assessments that resulted in the final model types and specifications for the DCS risk assessment algorithms that were presented in Chapter 5. The goal was to develop a set of models that improved upon the predictive accuracy of the existing set of algorithms in use by DCS. As noted previously, DCS requested that serious arrest or revocation be used as the outcome, where serious arrest was defined as a felony or misdemeanor violent arrest. The final set of nine models include static and dynamic factors and provide separate predictions for men and women by supervision type (straight probation, split probation, and parole) and three consecutive time periods: first quarter of supervision (Period 1), next three quarters of supervision in the first year (Period 2), and one-plus years of supervision (Period 3). The dynamic factors and period-specific models help account for the timing and occurrence of changes in risk over time.

The process of identifying the models that provided the best model fit and prediction included examinations and assessments of:

- Model specification (i.e., static and dynamic factors)
- Temporality (i.e., time)
- Model type (classification, survival, and ML).

This chapter begins by describing the identification of and tests for appropriate static and dynamic measures as well as determination of the appropriate time periods. Results from a selection of ML algorithms and survival models that were explored as possible alternatives to the logistic regression classification models are then presented.

This chapter then addresses the consideration of race and how we assessed the models for bias in predictive accuracy. Lastly, we describe a novel approach at incorporating uncertainty into predictions from our final models.

All models use the outcome of felony or violent misdemeanor arrest or revocation. As described previously, to assess model performance, we use the area under the ROC curve. The ROC curve plots the true positive rate compared to the false positive rate at every threshold. From the ROC curve, the AUC is calculated to provide a comprehensive, threshold-neutral statistic that represents the probability that the model correctly predicted the observed class (e.g., arrest) (Huang & Ling, 2005). AUC values can range from 0 to 1.0, with 0 being perfectly incorrect prediction, 0.5 akin to random chance (i.e., flipping a coin), and 1.0 being perfect prediction. Research evaluating criminal justice risk assessments identifies the AUC values associated with risk instruments as poor (0.50–0.54), fair (0.55–0.63), good (0.64–0.70), or excellent (0.71+) (Desmarais et al., 2017). These analyses were conducted with the pROC package in the R statistical software to produce tests of differences in paired AUC values (Robin et al., 2011).

6.1 Model Specification

Both static and dynamic factors were examined in the development of the models. Static factors are measured at the start of supervision and do not change over the course of supervision. Dynamic factors may change over time—either by switching on or off (e.g., employment) or by incrementing over time as an event reoccurs (e.g., count of positive drug tests during supervision).

6.1.1 Static Factors

Static factors included age measured at the start of supervision, race (used for model development but not for prediction), contextual measures of supervision at supervision start, and criminal history measures. (Sex and supervision type define strata in the dataset and are not factors in the models.) For individuals on probation, static factors include details drawn from the court case docket, such as the most serious underlying charge. This measure is operationalized as a categorical measure comparing violent, property, public order, or other charges to a reference category of drug offenses. The models for probation include flags for prior prison terms in the preceding 5 years, as well as a proxy for prior supervision.

For those on parole, we include measures derived from GDC that are tied to the relevant preceding prison term, such as whether the prison admission was for a prior revocation or if it was linked to a conviction for a property offense. Additionally, we include a count of prison disciplinary reports observed during that prison term and a report by GDC indicating whether the individual had been flagged for a new or existing mental health problem. As parole special conditions are usually applied to the entire length of a parole sentence, we include an aggregated measure that indicates whether the individual has any special condition for their parole term (e.g, drug testing or community service).

Criminal history is commonly used in risk assessments because past criminal history has been linked as a predictor of future arrest in many recidivism studies (Brame et al., 2003). Yet, research has also identified that the relationship between criminal history and future recidivism is dependent on time—specifically the amount of time that has

passed since the prior arrest or conviction (Kurlychek et al., 2006). To assess the utility of various criminal history measures, we tested for differences in predictive accuracy using measures for (1) lifetime arrest counts before the start of supervision, (2) lifetime arrests by charge type, and (3) arrests categorized by charge type and years before supervision start.[7] We also present the results of a LASSO logistic regression model for felony and misdemeanor violent arrest to assist in identifying the best lookback periods for criminal history measures.

Table 6-1 provides and compares the AUCs for the following models:

1. Base models that model the outcome using only age at supervision start, race (White or non-White), whether the individual was flagged as a potential gang member at the start of supervision, plus lifetime criminal history as a numeric count of arrests before the start of supervision
2. Base model plus lifetime arrests before supervision start by charge type, with charges categorized as violent, property, drug, public order, or probation/parole offenses
3. Base models plus lifetime arrests with limited lookback periods for arrest histories.[8]

The AUCs suggest that any of these models produce “fair” to “good” results, as all AUCs are above 0.55 and some are above 0.64. For men and women on straight

[7] As mentioned in the data description chapter, this analysis uses prior arrests as opposed to convictions because the completeness in reporting conviction data in the state of Georgia varies by the conviction county.
[8] Multiple lookback periods were assessed to identify the “best” periods, which were either a 2-year lookback period or a 5-year lookback period, depending on the model. LASSO logistic regression was used to assist in identifying the best lookback periods for criminal history measures. These results suggested little if any penalty from a limited lookback period (see Appendix B).

probation, including details on the charge types for prior arrests results in slight but significant improvement in AUCs. Moreover, limiting the lookback period for prior arrests (by charge) further improves AUCs in predicting serious arrest or revocation. This pattern is mirrored for men on split probation sentences, as adding detail on the type and timing of prior arrests results in improved model accuracy. We observe a similar, yet attenuated pattern for men supervised on parole. Although this general pattern is observed for women on split probation or parole, these differences in predictive accuracy are not significant at the $p < 0.001$ level. These results suggest that, when using only static criminal history, models of rearrest can be improved in many cases by including details on both the type of charge and the timing.

Table 6-1. Comparison of AUC Values for Static Models of Rearrest by Supervision Type, Sex, and Criminal History Type

Supervision Type	Sex	Base Model*	Lifetime Arrest by Charge Type**		Arrest by Charge Type and Lookback Period***	
		AUC	AUC	P-Value	AUC	P-Value
Straight Probation	Men	0.658	0.662	0.000	0.686	0.000
	Women	0.660	0.664	0.000	0.691	0.000
Split Probation	Men	0.633	0.640	0.001	0.658	0.000
	Women	0.609	0.613	0.363	0.628	0.218
Parole	Men	0.669	0.676	0.024	0.681	0.004
	Women	0.680	0.694	0.022	0.709	0.043

*Age, race, gang membership, and lifetime count of all arrests.
**Age, race, gang membership, and lifetime arrest count by charge type.
***Age, race, gang membership, and lifetime arrest count by charge type and limited lookback periods.

In addition to criminal history, the IDRACS models also include measures that describe the context of supervision at supervision start date. For individuals on probation, this includes details drawn from the court case docket, such as the most serious underlying

charge. This measure is operationalized as a categorical measure comparing violent, property, public order, or other charges to a reference category of drug offenses. The models for probation include flags for prior prison terms in the preceding 5 years as well as a proxy for prior supervision.

For those on parole, we include measures derived from GDC that are tied to the relevant preceding prison term, such as whether the prison admission was for a prior revocation or if it was linked to a conviction for a property offense. Additionally, we include a count of prison disciplinary reports observed during that prison term and a report by GDC for if the individual on parole had been flagged for a new or existing mental health problem. As parole special conditions are usually applied to the entire length of a parole sentence, we include an aggregated measure that indicates if the individual has any special condition for their parole term (e.g, drug testing, community service).

6.1.2 Dynamic Factors

Using longitudinal data to model recidivism has many benefits, including the ability to order events and to observe changes over time that might precede an arrest. Prior studies of recidivism during community supervision highlight the benefit of using dynamic features in improving predictive accuracy, especially when using proximal, rather than distal, changes (Brown et al., 2009; Greiner et al., 2014; Yukhnenko et al., 2020). However, some research has indicated that dynamic factors are not associated with increased predictive accuracy when modeling recidivism for individuals on supervision (Caudy et al., 2013). This section describes the dynamic features included in the IDRACS model and reports the results of tests of predictive accuracy.

As many of the dynamic measures are collected because of regular supervision practices, nearly all of these factors reflect changes observed while someone is being supervised through regular or intensive supervision. These factors include positive and negative drug tests, operationalized as a running count throughout supervision; and technical violations, aggregated by type and listed as a running count during supervision. Other measures include a flag indicating that the individual was employed (available only for the first year), a warrant flag indicating that DCS has received notice of an outstanding warrant for the individual, and a contact flag indicating that the individual's supervision level has been reduced to contact. Other dynamic counts are non-violent misdemeanor arrests observed during supervision by type (drug, public order, property, probation/parole offenses) and the number of unique supervision-level changes (e.g., high to standard, standard to contact). In addition, after the first 90 days, a measure indicating whether someone has or has not been tested for drugs in the previous 90 days (regardless of the outcome of the test) was included. Finally, for those on probation, special conditions of probation were derived from probation dockets to identify special conditions assigned to the probation term, including flags for community service, drug/alcohol treatment or testing, educational or employment requirements, fees, no contact orders, violent behavior treatment, or other conditions. As the special conditions are tied to a docket, they can end before the end of the supervision term and are thus included as dynamic factors.

6.1.3 Comparisons of Predictive Accuracy of Static vs. Dynamic Models

Table 6-2 shows the differences in AUC values obtained by supervision type and sex for models that feature only static vs. static and dynamic features. The measures in these

models are not time-specific (as described later), because they only show changes observed after supervision start, but before the last day the individual is observed in the dataset. As is evident, all models are substantially improved by including both static and dynamic features, with increases in AUC ranging from 0.15 to 0.19, which represent substantive as well as statistically significant gains in accuracy. Per the Desmarais et al. (2017) classification, these models that include dynamic factors would be classified as providing excellent predictive fit.

Table 6-2. Comparison of AUC for Models of Rearrest Featuring Static or Static/Dynamic Features by Sex and Supervision Type

Supervision Type	Sex	Static Model	Dynamic Features		
		AUC	AUC	P-Value	Difference
Straight probation	Male	0.671	0.836	0.000	0.165
	Female	0.670	0.837	0.000	0.166
Split probation	Male	0.647	0.818	0.000	0.172
	Female	0.627	0.820	0.000	0.193
Parole	Male	0.668	0.839	0.000	0.171
	Female	0.720	0.872	0.000	0.152

Source: Test AUC value comparisons from longitudinal cohort of individuals on probation and parole in Georgia, 2016–2019

6.2 Identifying Distinct Time Periods of Risk of Rearrest/Revocation

The previous classification model results use a maximum 4-year period of risk exposure, as the data range from January 1, 2016, to a study cutoff period of December 31, 2019. However, exploration of the data via survival analyses shows that individual risk of reoffending peaks in the beginning of supervision. This study explored using predicted hazard rates derived from survival models, but the format of this output makes employing predictions from survival models difficult to implement. To account for

variation in risk over time, while still using a classification framework, we tested multiple different periodized logistic regression models. Tests included monthly and quarterly models, with aggregations of time periods aimed at maximizing predictive accuracy in the resulting models. These tests revealed three distinct time periods: (1) the first quarter year of supervision (Period 1), (2) the remaining three quarters of the first year (Period 2), and (3) a post-one-year supervision period (Period 3). In addition to providing useful predictions, these periods also mirrored supervision practices in Georgia, where the first 90 days comprise intensive supervision along with reentry programming. The remaining first year of supervision often consists of standard supervision, with in-person check-ins. However, after one year of supervision, many individuals on probation who have not been revoked and who do not remain on intensive supervision can be transitioned to “contact” supervision, which requires telephonic check-ins with a call center.

We compared models developed for specific time periods to models developed using the full data range or *Single Period* models. As a reminder, the Period 1, Period 2, and Period 3 models are estimated using only individuals who entered each period, as those experiencing the outcome or leaving probation for any other reason are dropped from the data in successive periods. Thus, only the Period 1 models are estimated using the same data as the Single Period models.

We first present the results of unpaired statistical tests of differences in two AUC values, comparing the accuracy of the predictions from the Single Period model and the accuracy of the predictions for models that use a limited period (i.e., Period 1, Period 2, and Period 3). These tests are unpaired because we are comparing predictive accuracy

between two models that have different sample sizes, and, in this analysis, are derived from different time periods. They are less useful compared to paired tests but can provide insight if one model substantially outperforms the other. We then turn to an analysis of paired tests of AUC values that compare AUC measures for the time-specific models (e.g., Period 2) against the AUC value obtained when using the Single Period model to predict the outcome in that period (e.g., using the Single Period model to predict the outcome for the Period 2 dataset).

Table 6-3 shows the AUC for each of the four models (Single Period, Period 1, Period 2, and Period 3) for each supervision type for men and women, and p-values for the unpaired statistical tests comparing the Single Period model AUC to the AUC for the other models. Of interest is whether a model has a higher AUC value between the pairs and the difference is significant at $p < 0.001$ (our established significance level because of the large sample sizes). In only two cases of 18 tests does the Single Period model outperform the time-specific models—the Single Period model is "better" than the Period 1 model for the models for men on straight and split probation.

Table 6-3. Unpaired Tests Comparing Time-Specific and Single Period Models

Supervision Type	Sex		Single Period Model*	Period 1*	Period 2*		Period 3*	
		AUC	AUC	P-Value**	AUC	P-Value**	AUC	P-Value**
Straight probation	Men	0.836	0.801	0.000	0.825	0.068	0.828	0.231
	Women	0.837	0.801	0.001	0.851	0.116	0.855	0.110
Split probation	Men	0.818	0.748	0.000	0.795	0.013	0.789	0.006
	Women	0.820	0.759	0.056	0.816	0.878	0.833	0.614
Parole	Men	0.839	0.812	0.109	0.858	0.128	0.819	0.240
	Women	0.872	0.731	0.027	0.901	0.314	0.813	0.253

*Single Period = all observed time; Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.
**P-value of unpaired statistical test comparing the AUC for the period to the AUC for the Single Period model.

Table 6-4 provides the results for the paired tests in which the AUCs from the period-specific models are compared with the AUCs obtained when using the Single Period model to predict outcomes within Period 1, Period 2, and Period 3. Here, we see that the period-specific models outperform the Single Period model in most cases. In particular, the period-specific models are better at predicting rearrest or revocation in the first year of supervision (usually before individuals are transitioned to contact status). The exceptions are the models for (1) women on parole (all periods) and women on straight and split probation in Period 3, where differences are not significant at the 0.001 level; and (2) men on straight and split probation and on parole in Period 3, where the Single Period model performs significantly better. However, this finding might be expected, as the Single Period model uses data from the last day observed in the dataset, which for many will be beyond one year. These findings suggest that period-specific models may be useful and provide more accurate predictions in real-world settings, where the risk of arrest may vary depending on one's time on supervision.

Table 6-4. Paired Tests Comparing Time-Specific and Single Period Models

Supervision Type	Sex	Period 1 Data			Period 2 Data			Period 3 Data		
		Period 1 Model	Single Period Model		Period 2 Model	Single Period Model		Period 3 Model	Single Period Model	
		AUC	AUC	P-Value	AUC	AUC	P-Value	AUC	AUC	P-Value
Straight probation	Men	0.801	0.611	0.000	0.825	0.684	0.000	0.828	0.860	0.000
	Women	0.801	0.662	0.000	0.851	0.754	0.000	0.855	0.849	0.329
Split probation	Men	0.748	0.566	0.000	0.795	0.650	0.000	0.789	0.862	0.000
	Women	0.759	0.588	0.000	0.816	0.669	0.000	0.833	0.871	0.018

Supervision Type	Sex	Period 1 Data			Period 2 Data			Period 3 Data		
		Period 1 Model	Single Period Model		Period 2 Model	Single Period Model		Period 3 Model	Single Period Model	
		AUC	AUC	P-Value	AUC	AUC	P-Value	AUC	AUC	P-Value
Parole	Men	0.812	0.677	0.000	0.858	0.763	0.000	0.819	0.867	0.000
	Women	0.731	0.784	0.445	0.901	0.829	0.028	0.813	0.863	0.147

*Single Period = all observed time; Period 1 = first quarter, Period 2 = next three quarters of first year, Period 3 = one-plus years of supervision.
**P-value of paired statistical test comparing the AUC for the period to the AUC for the Single Period model using data for the specific period.

6.3 Assessing Racial Bias in Prediction

In addition to developing a series of models that predict accurately, this project also sought to build an unbiased predictive tool. Although risk assessments can provide improvements, they run the risk of exacerbating existing biases in the data, specifically racial bias in prediction. Recent studies have examined racial biases in RAIs and predictive models with the goal of ensuring that risk scores predict similarly regardless of the race of the individual (Berk et al., 2018; Skeem & Lowenkamp, 2016). To counter potential biases, we employ a “bias-aware” approach to modeling that specifically models inherent group differences and attempts to extract the contribution of group variables (e.g., race) from the prediction (Cardoso et al., 2019). Using this framework can result in lower model accuracy as a tradeoff for ensuring that predictive techniques do not exacerbate inherent biases. In practice, when developing the predictive models in the training datasets, we include race as a dichotomous indicator (White/non-White) in the training dataset. However, for prediction we omit this predictor from our linear model (essentially, multiplying the coefficient by 0 for every individual).

To assess the presence of predictive bias, we compare for predictive accuracy to ensure that our models are not predicting differently by race. When examining differences, we use the period-specific models to predict racially stratified test datasets. We then compare these AUC values using unpaired statistical tests of the differences in AUC values to determine if a model's predictions are more or less accurate depending on the race of the individual.

As seen in Table 6-5, differences are minimal in the predictive accuracy for models when predicting for White vs. non-White populations. In the Period 1 models, predictions for non-White individuals are mostly more accurate, but the differences with the predictions for White individuals do not reach significance, indicating there is no difference in the AUC values and that the models do not perform differently from each other. This pattern persists across supervision type, sex, and period. Furthermore, only one model has a difference in AUC value that reaches the $p < 0.001$ level (the Period 1 model for women on parole). In this instance, the model predicts significantly better for non-White than White individuals. However, in this model, only two non-White females were arrested; thus, the test dataset for non-White individuals did not contain sufficient variation to generate reasonable inference. However, the lack of sufficient sample size for females on parole in the test dataset is a limitation of the study driven by low overall numbers of women on parole in Georgia compared to men or women supervised on straight or split probation.

Table 6-5. Comparison of Predictive Accuracy by Race (White and Non-White) for Time-Specific Models by Supervision Type and Sex

	Sex	Period 1			Period 2			Period 3		
		White	Non-White		White	Non-White		White	Non-White	
		AUC	AUC	P-Value	AUC	AUC	P-Value	AUC	AUC	P-Value
Straight probation	Men	0.831	0.824	0.522	0.834	0.817	0.082	0.806	0.797	0.406
	Women	0.850	0.860	0.608	0.844	0.849	0.766	0.794	0.806	0.579
Split probation	Men	0.816	0.774	0.024	0.813	0.784	0.058	0.754	0.742	0.584
	Women	0.770	0.744	0.657	0.832	0.785	0.270	0.820	0.851	0.460
Parole	Men	0.812	0.813	0.986	0.849	0.866	0.337	0.845	0.804	0.178
	Women	0.691	0.957	0.000	0.888	0.936	0.212	0.789	0.878	0.246

Note: Unpaired test of AUC value comparisons.
Tests are produced using the pROC package in R.

6.4 Machine Learning Model Investigation

ML classification models often yield substantial gains in accuracy when compared to traditional inferential models. However, these models have limitations in interpretability, notably that identifying the precise nature of the relationships between the outcome and the predictor variables can be difficult. To gauge whether substantive gains in accuracy can be achieved by using ML models when compared to traditional inferential models (i.e., logistic regression), we compared the AUC values when using a random forest classifier and a GBM classifier to the AUC values obtained when using a logistic regression model.

Given the large sample sizes involved in our training and test data, observing a significant difference in a paired test of AUC values is not out of the ordinary, even for minor differences in actual values. To account for this, we set two qualifying standards:

1. We require a p-value of < 0.001 from the tests of paired AUC values to indicate the presence of a statistically significant difference in AUCs.

2. We require an observed difference in AUC values of 0.02 to indicate that a particular classifier is producing substantive gains in accuracy compared to the model it is being compared to.

Table 6-6 highlights several findings. First, the random forest models usually slightly outperform the logistic regression models. Similarly, the GBM models also slightly outperform the logistic regression models, but less so when compared to the random forest classifiers. However, when examining the differences in the AUCs, we rarely observe a difference of more than 0.02 in the AUCs, which would represent a small increase in accuracy. Notably, only three models, all in the parole supervision type, result in the ML models outperforming the traditional logistic regression models and producing differences in AUC values that are equal to or greater than 0.02. In addition, there are relatively few differences where the p-value is < 0.001, indicating a statistically significant difference in model performance, and all these differences are observed in the Period 2 models.

Table 6-6. Comparison of Logistic Regression Model to Random Forest and GBM Classifiers

Model Type	Parole	Straight Probation	Split Probation
Lifelines Cox Survival Models			
Concordance*	0.885	0.836	0.827
Gradient-Boosted Survival Models			
Concordance*	0.871	0.843	0.823
Brier Score	0.062	0.117	0.124

Model Type	Parole	Straight Probation	Split Probation
Random Forest Survival Models			
Concordance*	0.863	0.852	0.855
Brier Score	0.117	0.104	0.099
Deep Survival Models			
Concordance*	[0.498, 0.548, 0.490]	[0.077, 0.848, 0.813]	[0.854, 0.82, 0.828]
Brier Score	[0.093, 0.151, 0.205]	[0.077, 0.117, 0.125]	[0.071, 0.115, 0.157]
Recurrent Deep Survival Models			
Concordance*	[0.8560, 0.743 0.690]	[0.870, 0.855, 0.792]	[0.827, 0.822, 0.781]
Brier Score	[0.015, 0.036, 0.063]	[0.076, 0.119, 0.136]	[0.088, 0.139, 0.147]

Concordance is shown as comprehensive concordance for the Lifelines Cox model and both tree-based survival models. However, for the deep learning survival models, concordance is presented in the 3 distinct time periods (thirds of the data).

Table 6-8 presents the results but for men by supervision type. Like the results for the women's models, there is little difference in predictive accuracy for men on parole. There is slight improvement in concordance for males on straight and split probation with the random forest models outperforming both the gradient-boosted and Cox models.

Table 6-7. Concordance and Brier Scores for Cox and ML Survival Models for Men on Supervision

	Parole	Straight Probation	Split Probation
Lifelines Cox Survival Models			
Concordance*	0.859	0.829	0.802
Gradient-Boosted Survival Models			
Concordance*	0.855	0.848	0.829
Brier Score	0.096	0.117	0.129
Random Forest Survival Models			
Concordance*	0.867	0.867	0.847
Brier Score	0.104	0.108	0.12
Deep Survival Models			
Concordance*	[0.850, 0.839 0.845]	[0.870, 0.855, 0.792]	[0.827, 0.822, 0.781]
Brier Score	[0.0502, 0.121, 0.15]	[0.076, 0.119, 0.136]	[0.088, 0.139, 0.147]

	Parole	Straight Probation	Split Probation
Recurrent Deep Survival Models			
Concordance*	[0.7953, 0.716 0.705]	[0.870, 0.855, 0.792]	[0.827, 0.822, 0.781]
Brier Score	[0.0168, 0.0398, 0.071]	[0.076, 0.119, 0.136]	[0.088, 0.139, 0.147]

Concordance is shown as comprehensive concordance for the Lifelines Cox model and both tree-based survival models. However, for the deep learning survival models, concordance is presented in the three distinct time periods (thirds of the data).

6.5 Incorporating Uncertainty into Predictions

Beyond identifying the most accurate, useful, and parsimonious model types and model specifications, this project sought to incorporate the concept of uncertainty in the predictions derived from these models. Understanding the uncertainty associated with a prediction is an integral step in making use of the prediction, as not all predictions from the same model are equally accurate or confident. Knowing the limitations of a prediction is also key when making operational decisions based on the prediction, as agencies have limited resources to assign to intensive supervision and can benefit from focusing on the most certain and accurate high-risk predictions. To incorporate uncertainty into the output of our predictions, we employed a bootstrapped prediction method that derives the classification models described above from 1,000 subsets of the model-specific training data to generate an average prediction (the mean of these predictions) and a range in the predictions (the minimum and maximum observed predicted probabilities), as described below.

Once a logistic regression model is fit to training data, the predicted probability of the rearrest outcome can be computed for each individual in either the training data or a test dataset. However, estimating the standard error of the predicted probability is important, as illustrated using a simple example. If two individuals in a test dataset have

the same predicted probability of rearrest (e.g., 0.7), but very different predictors, we may want to know whether the precision of the 0.7 estimate is the same for both individuals. An estimate of this precision is given via the standard error of the predicted probability of rearrest. Suppose that Individual A has average values on all the predictors, while Individual B has extreme values for all the predictors. Under this situation we would expect the standard error of Individual A's predicted probability of rearrest to be smaller than the standard error for Individual B.

The usual approach to estimating the standard error of a predicted probability is the delta method approach (see Agresti, 2012; Xu & Long, 2005). If one is using software where this method is already implemented (e.g., in the predict.glm function in R; Baldwin et al., 2010), this approach is fast, and the resulting standard errors can be used to calculate confidence intervals for each individual's predicted probability of the arrest/revocation outcome. In contrast, most software for developing a dashboard for displaying confidence intervals for predicted probabilities does not have the delta method standard errors implemented, although these calculations can be derived manually. An alternative to the complex implementation of the delta method standard errors was to bootstrap the standard errors using the process explained below.

The process for bootstrapping cases used the Boot function of the R (Baldwin et al., 2010) package car (Fox & Weisberg, 2019) to get 1,000 samples of regression coefficients. This process was repeated for each period, sex (male, female), and supervision type (straight probation, split probation, parole). For any individual, the bootstrap estimate of the standard error of their predicted probability is the standard deviation of the 1,000 bootstrapped estimates of their predicted probability. These

bootstrapped standard errors were found to be identical to the delta method standard errors within rounding to three to four decimal places. The added benefit is that the bootstrapping method can be applied to a more general set of machine learning models, given that it derives its range by producing predictions from a sub-sample of the data. Additionally, bootstrapping provides a broader picture of the uncertainty, as researchers are able to observe the distribution of predictions as opposed to just the range which is observable from the delta method.

To continue the hypothetical example where two individuals had a predicted probability of 0.7, their standard errors would be used to calculate 95% confidence intervals—for example, (0.6, 0.8) for Individual A and (0.35, 0.95) for Individual B. In this situation, we can be highly confident that Individual A has a greater than 50% chance of rearrest, while Individual B has a great deal of uncertainty in their predicted probability. As such, a probation or parole officer might approach case planning differently for these two cases.

Figure 6-1. Example of Risk Scores with Confidence Intervals

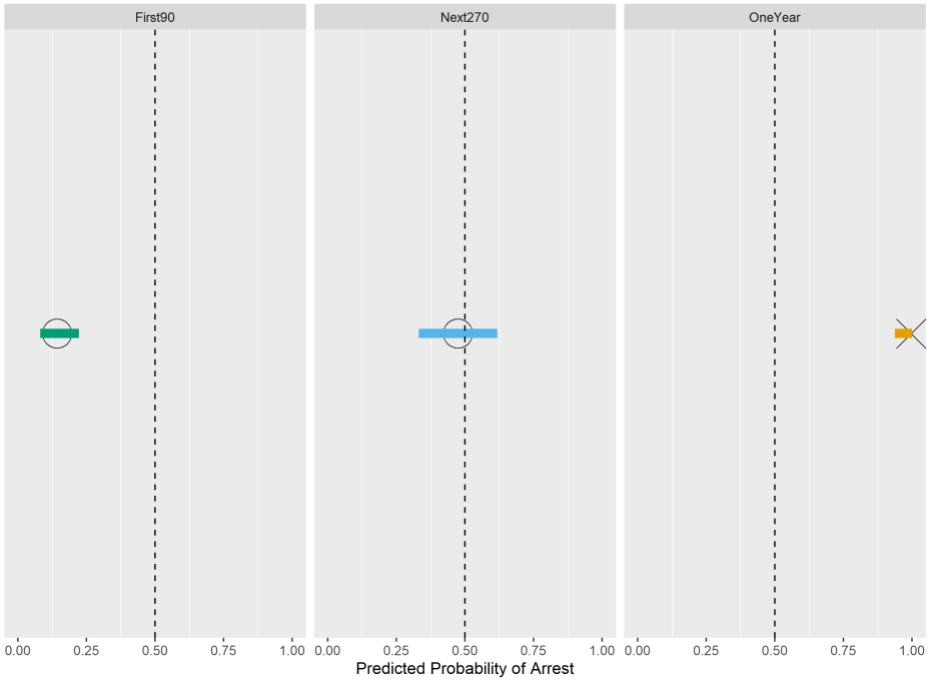


Figure 6-1 illustrates the potential for inclusion of an error estimate for prediction. The figure illustrates three potential risk scores for an individual during the three supervision periods. In Period 1, the individual is scored low (point estimate of about 0.10) with a relatively small confidence interval. Because of changes in dynamic variables as the individual moves into Period 2, the Period 2 model predicts a greater likelihood of arrest/revocation with a point estimate of about 0.5, but a much larger confidence interval. The circles around the first two points indicate that the individual was not arrested/revoked during Period 1 or 2. After the first year of supervision, again because of changes in dynamic factors, the individual's risk score approaches 1 with very little uncertainty; and, in fact, the individual was arrested.

Given that the predicted probabilities from risk scores are not generated and used in a vacuum, and probabilities are often compared to each other, incorporating uncertainty into these estimates can be valuable when making operational decisions. If percentile or point-based thresholds are used to assign services, supervision levels, or supervision

strategies, understanding that scores around these thresholds might be wide-ranging can be useful for allocation. However, the full implications of incorporating uncertainty into predicted probabilities, especially for operational decisions, has yet to be explored.

6.6 Analytical Results Summary

This chapter described the results of an extensive and exploratory modeling process. Although including all the findings from this practice is impractical, several key takeaways resulted from the model development process. First, time- and type-specific criminal history appear to be more useful for prediction compared to broad and comprehensive criminal history measures. Building on this, when prioritizing predictive accuracy, using a limited recall period for criminal history to incorporate a "decay effect" of lifetime criminal history is justified. In addition, using data collected during community supervision (i.e., dynamic measures) compared to using only static measures collected during or before intake results in substantial gains in model accuracy. Correspondingly, employing these dynamic features can include both protective factors (e.g., negative drug test results, recorded employment) or risk factors (e.g., misdemeanor arrests, technical violations, positive drug tests), and changes in these factors can lead to both an increase *and* decrease in an individual's predicted probability of rearrest.

Beyond feature identification, this study also provided support for examining the risk of rearrest as specific to certain periods of time. Using a survival analysis-driven perspective that is built into a classification framework (i.e., creating time-specific classification models) proves useful in predicting rearrest and accounts for the fact that the nature of rearrest often changes throughout the course of supervision. In this study, we identified that the first 90 days of supervision were distinct from the remaining three

quarters of the first year and beyond, and that these distinctions were useful for prediction. Importantly, testing the utility of ML algorithms (e.g., random forest or GBM) yielded minimal or no gains in predictive accuracy compared to traditional inferential statistical models such as logistic regression. Lastly, this study explored the practical application of incorporating uncertainty into the resulting predicted probabilities of rearrest.

7. Model Integration

An important part of this project's aim was to not only develop a useful tool for predicting rearrest and revocation for DCS, but to assist in the integration process for the tool to ensure that the practical application of this study (i.e., the IDRACS risk algorithms) could be incorporated into DCS's operations. Integrating the IDRACS tool into DCS's CMS included two main tasks: (1) data management and (2) model application.

In addition to providing documentation and materials, we describe the process for quality assurance, necessary to ensure that the IDRACS tool was implemented as designed. RTI, in partnership with our collaborators Applied Research Services, Inc. and DCS, approached this task in a collaborative manner, with members of DCS information technology (IT) staff assigned to translate the data management and modeling code before implementation. RTI staff provided documentation and pseudo syntax. Furthermore, both parties agreed on a series of comparison tasks using DCS data to ensure that the data management and modeling code were being translated and implemented to produce expected results. This chapter describes this process.

Gaining insight from CSOs and their supervisors was an important part of the process. Focus groups were held with CSOs and management early in the development process. Results of the IDRACS project were presented and discussed with CSOs as part of the implementation process. This chapter concludes with a summary of their thoughts.

7.1 Data Management

Ensuring accurate data processing is a crucial step in tool implementation. A substantial proportion of the study was spent during the development of the IDRACS tool in data

management to (1) understand the data and (2) produce useful measures that could be used for prediction. However, a failure to accurately translate code that was developed for a research longitudinal database to a live production database could result in unexpected and inaccurate results when attempting to generate model predictions. This section describes the process of translating RTI's code to produce equivalent variables and comparing data management results on test databases to verify accurate translation.

7.1.1 Adapting Longitudinal Data Management to One-Day Measures

To assist DCS with integrating the final set of risk models, the original longitudinal data management code required adaptation to align with how the measures are calculated in a production setting. For production, the calculation of the static measures, as of the start of supervision date, remain similar to their longitudinal counterparts, while dynamic factors were reworked to be calculated between the start of the supervision period and a set date agnostic of end dates that are only available in historic datasets. The approach for how to accomplish adapting the existing longitudinal code was developed in partnership with DCS to (1) provide them with example code so that they could derive and implement creation of the measures internally, and (2) have a comparable dataset for evaluating their implementation. To carry out this process, RTI produced a series of Jupyter notebooks in Python, which combined code and plain text language to walk through the measure derivation process.

The IDRACS models include a series of static variables that are calculated as of the start of the supervision period and, in the original longitudinal dataset, did not change. The concern raised was that the risk models were created to predict felony or violent

misdemeanor arrests or revocation, at which point individuals were truncated from the modeling cohort. In practice, DCS indicated that a proportion of individuals who experience the truncating event continue on supervision with the same supervision episode (i.e., the same probation term drawn from a court docket) despite being detained in county facilities or in probation detention centers. The practical solution for this issue, as agreed upon by DCS, was that if an individual experienced a felony arrest or violent misdemeanor arrest during their supervision episode but was identified to continue on active supervision, the “days since supervision start,” as used for model assignment, would be reset for the purposes of calculating risk scores. This reset process impacted the creation of static variables by providing a new date from which the supervision start (for modeling purposes) would be calculated and ensured that the outcome they experienced would be reflected in the updated criminal history information.

The dynamic variables that are used in the IDRACS risk model also required translation when moving from data management used in a historical longitudinal dataset to the single-day measures necessary for a production database. This is largely because in a production setting, the supervision episode does not have an end date to use to filter only to events of interest. In practice, the end date becomes the date when the measures are being calculated (i.e., “today”), and for the counting measures, the data are filtered to any events of interest that occurred between the start of the supervision period and the current date. For events that have start and end dates of their own, such as employment—which, in the risk models, are represented as binary indicators—the variable syntax was adapted to handle missing end dates both for the supervision

period and the event of interest. For example, when the start date for the employment event is before the current date and the employment event is still open, the employment measure is coded as a 1 in the single-day measures, whereas in the longitudinal dataset the start and end dates for both the supervision period and the employment event are compared to identify when the indicator is turned on by being set to 1, and then turned off or set back to 0.

7.1.2 Data Comparison Process

After providing DCS with the notebooks containing the syntax used to create the measures included in the IDRACS algorithms, modeling staff developed a process for comparing the measures DCS derived against the versions derived by the model team. DCS provided a test database (i.e., a dataset with active cases that was frozen at a certain time point) used for integration, which contained their production data frozen on 09/12/2020. This test database overlapped substantially with the research data extract that RTI used to develop the IDRACS models. This dataset was used as a starting point to create a comparison cohort of individuals who started probation or parole between 2016 and 2020 and were available to both teams. This dataset was used to calculate the required measures as of 09/12/2020. Each team independently calculated the measures to compare output and adjudicate any differences. Small differences were expected because of DCS using the frozen production database and RTI using a static cut of the production database from a later date, but larger differences were investigated and addressed by DCS and RTI.

7.1.3 Measure Reconciliation

During the integration and data comparison process, RTI and DCS held weekly discussions between RTI project staff and the DCS integration team, which included members of the IT and operations staff at DCS. The diverse set of opinions led to the revision of how certain measures are defined, work in reconciliation to ensure that measures were the same, and the refinement of model specifications to incorporate DCS feedback.

One example of this reconciliation was the decision to make the measures derived from court dockets (e.g., the indication of split sentence and the worst offense from the attached dockets) dynamic instead of static. While RTI developed the models from longitudinal datasets using a subset of active dockets, individuals with probation sentences can have multiple dockets added or end during the contiguous period of supervision. Originally, RTI developed the split sentence indicator and other docket fields to limit to only dockets within a certain window around the start of supervision, given the potential for data errors. However, operationally, DCS wanted to ensure that new or dropped dockets during the supervision episode were reflected in an individual's risk score.

In addition, DCS confirmed that drug testing after the initial supervision period is done for cause and is discretionary. To account for the cessation of drug testing, DCS suggested an additional drug testing measure for individuals who have not been drug tested in the prior 90 days, since this is an indication that the supervising officer does not believe it to be necessary. RTI derived and tested this measure in training and test

datasets and confirmed that it improved model accuracy and incorporated the measure in models after the first 90 days of supervision.

Identifying special probation conditions also presented a unique challenge during the original derivation of the IDRACS measures for model development, and proved difficult to reconcile during the integration process. This difficulty is because probation special conditions are captured from open text fields listed in the court docket. As such, deriving the categorical measures required running keyword searches, which inherently leads to over- and under-counting of conditions in all categories. Different keyword search methods, such as regular expressions or whole word matching, lead to prioritizing false positives or false negatives, respectively. After initial comparisons yielded substantial differences in the measure (a product of different approaches to the keyword search), the integration team agreed to prioritize false negatives over false positives, since false positives could raise an individual's risk score erroneously.

7.2 Model Application

Once both parties agreed on the generation process for the measures and ensured that the comparison datasets were equivalent, the integration process transitioned to ensuring that DCS could use RTI's trained models to predict scores daily for the people under supervision. This process required less reconciliation, given the efforts dedicated to ensuring that data management was comparable.

RTI used logistic regression for its modeling method, which (as an extension of linear models) creates a vector of coefficients, each element corresponding to one measure along with an intercept. To produce a distribution of predictions that included uncertainty in the predictions, as discussed in Chapter 6, 1,000 versions of the model were trained

and saved, allowing each measure's coefficients to vary according to their uncertainty. When applied to a single person, this generated 1,000 predictions for that person, which could be turned into a mean prediction and a corresponding confidence interval. Equation 1 shows the translation of one model's coefficients and values used to produce a predicted probability:

$$probability_p = \frac{e^{(measures_p \cdot coefficients_j)}}{1 + e^{(measures_p \cdot coefficients_j)}}$$

Equation 1: The probability of person p given coefficient set j

The benefit of this straightforward approach is that given a set of input data and the coefficients, the predictions for each person would be deterministic based on the implementation of a discrete set of mathematical operations. To isolate any errors in this step from any differences in the data, both RTI and DCS predicted on one dataset, namely the measures calculated by DCS for each person in their test database. In practice, RTI generated predictions on the test dataset provided by DCS, which agency IT staff then confirmed by replicating the same predictions on their end.

7.3 Focus Group Testing and Officer Perceptions

As part of the integration task, RTI and Applied Research Services carried out a series of focus groups conducted at DCS field offices to solicit feedback from CSOs on the use and impact of risk assessment, and to identify questions officers may have during implementation. Six focus groups were conducted during the weeks of November 13 and December 11, 2023. These focus groups were conducted in three urban sites, two rural sites, and one suburban site in north, south, and central areas of Georgia. Staff at the DCS central office aided in scheduling the focus groups, and local supervisors

selected at least five officers per site to participate in each focus group, resulting in a total of 26 officers.

The focus groups lasted approximately one hour. The facilitator provided a brief overview of the new instrument and asked questions about how officers currently use the risk score, data they would like to see related to risk assessment, and ideas to make the new tool user-friendly for the field. Feedback was solicited by showing officers actual cases with risk scoring computed by the current and new risk instruments. As part of the focus groups, officers provided perspective to potentially improve implementation statewide. The primary insights from these discussions include the following:

1. Officers do not currently use the risk score to prioritize their caseload and instead use a combination of supervision level (e.g., standard, high, specialized, or contact) and identified needs. For officers, the risk score was primarily a tool used to determine supervision levels and was only referenced if requesting a supervision-level override.
2. When shown test cases, most officers preferred the results of the new instrument over the current instrument; many commented that it seemed to pick up on case nuances better.
3. Most officers wanted to know the factors that led a person to move from one supervision level to another. They thought this would help them identify ways to offer support and services, as well as to provide individuals with feedback on how to succeed (i.e., Your risk level increased because you have not been employed for 6 months—how can I support you in finding employment?).

These findings provide integral feedback about how line officers are likely to interact with the tool and what information would be useful to improve supervision upon implementation. In addition, these findings demonstrate the need for additional research on how line officers and other correctional staff may interact with automated systems and how these actions may impact operations or supervision practices.

7.4 Integration Summary

This chapter describes the process of translating algorithms derived from a research project into a production tool. The key concerns in integrating a predictive algorithm are ensuring equivalent data management and predictions. These issues stem from the fact that algorithms are often developed on extensively cleaned datasets, while operational predictions may use full databases with noisy and missing data. RTI worked with DCS IT staff to establish comparable data management practices in a test dataset environment, confirming measure creation and prediction before implementing the tool. Through this collaborative process, RTI and DCS verified the variable syntax and adjusted measurement to make sure these measures both (1) mirrored those used in algorithm development and (2) accounted for the differences between a longitudinal research dataset and an operational CMS database. This process, while labor-intensive, is crucial to verifying that predictive algorithms are integrated in a way that will result in useful and accurate predictions.

8. Model Revalidation and the Impact of COVID-19

Model revalidation is an integral step in predictive modeling, as changes in the underlying data can have substantial impacts on model accuracy and usefulness. Changes in the prevalence of the outcome (here, arrest), the prevalence of predictors, or the relationship between predictors and outcomes can all impact model performance. Revalidating risk instruments is common practice in criminal justice settings to ensure that predictive tools remain accurate (Cohen & Lowenkamp, 2018; Steadman et al., 2007).

The COVID-19 pandemic, which changed both conditions for crimes and legal system practices, had a substantial and often differential impact on crime rates throughout the United States (Boman & Gallupe, 2020; Hodgkinson & Andresen, 2020; Jahn et al., 2022). To address the potential impact of the COVID-19 pandemic on the utility of the IDRACS tool, we conduct a limited revalidation to explore how the prevalence of arrest and the model accuracy changed both during and after the COVID-19 pandemic. For these analyses, we define the “during COVID” period as March 1, 2020, to June 30, 2021, which coincides with widespread availability of the first COVID-19 vaccines. The post-COVID period begins on July 1, 2021, and extends to September 30, 2022.

The revalidation looks at new probation or parole starts during both periods. Importantly, using new starts caps the number of individuals who can progress through the first year of supervision to the Period 3 model. In the during COVID period, only new starts that occurred from March 2020 to June 2020 have the opportunity to progress through the models to the Period 3 model, and individuals who started probation on March 1, 2020, would have a maximum of 4 months of observation in the post-one-year period. For the

post-COVID period, we observe arrests through September 2022 but stop new starts in July 2022, to allow for a minimum of 90 days of exposure for individuals included in the model. Like the during COVID period, only individuals who started probation or parole between July and September 2021 have the opportunity to progress through supervision to make it to the Period 3 model. Again, for individuals who started supervision on July 1, 2021, they have an observation period of a maximum of 3 months through September 30, 2022. As such, given the differences in observation times between the pre- and during/post-COVID periods for the Period 3 model, there is limited utility for revalidation for this period.

The limitation in the inference for revalidating also applies to the Period 2 models, as the potential observation periods are naturally truncated for these groups. For the during COVID cohort, since the observation period stops at July 2021, we only have 5 months of new starts (those starting supervision between March 2020 and July 2020) where we observe at least one year's worth of post-start data. Similarly, for the post-COVID cohort, since the observation period is paused at September 2022, only individuals who had new starts on probation or parole from July 2021 to October 2021 (4 months of new starts) would have had 9 months of observation time after progressing through the first 90 days of supervision. Given the lack of comparable observation periods, the during COVID analyses for the Period 2 and Period 3 models are exploratory only. In addition, for the post-COVID analyses, more time will need to pass to sufficiently assess model performance for these periods.

8.1 During COVID Model Exploration

To examine how changes in criminal behavior and legal system practices may have affected the accuracy of the IDRACS model, we first explore differences in pre- and during COVID arrest rates before turning to comparisons of model accuracy in the pre- and during COVID periods. Table 8-1 shows the prevalence of felony or violent arrest by supervision type, sex, and modeling period, including the difference in arrest rates and the p-values for the statistical tests of differences in proportions. Surprisingly, difference is limited in the prevalence of arrest in the first period of supervision, regardless of supervision type or sex. However, as expected given the difference in observation periods, significant differences begin to emerge during Period 2. In these models, differences in arrest prevalence range from 4% (e.g., straight probation) to 7% (parole). As expected, differences are stark in prevalence rates in the Period 3 models, given the lack of observation periods of comparable length.

Table 8-1. Prevalence of Felony or Violent Arrest for the Before and During COVID Periods

Supervision Type	Sex	Model	Felony or Misdemeanor Violent Arrest		Difference	P-Value
			Before COVID	During COVID		
Parole	Women	Period 1	7%	5%	1%	0.334
		Period 2	16%	9%	7%	< 0.0001
		Period 3	13%	7%	6%	0.262
	Men	Period 1	9%	8%	1%	0.01
		Period 2	22%	16%	6%	< 0.0001
		Period 3	23%	10%	14%	< 0.0001

Supervision Type	Sex	Model	Felony or Misdemeanor Violent Arrest		Difference	P-Value
			Before COVID	During COVID		
Split probation	Women	Period 1	10%	8%	2%	0.051
		Period 2	21%	14%	6%	0.002
		Period 3	25%	7%	18%	< 0.0001
	Men	Period 1	9%	7%	2%	< 0.0001
		Period 2	24%	17%	6%	< 0.0001
		Period 3	32%	12%	20%	< 0.0001
Straight probation	Women	Period 1	12%	13%	-1%	0.045
		Period 2	18%	15%	4%	< 0.0001
		Period 3	19%	7%	12%	< 0.0001
	Men	Period 1	13%	16%	-3%	< 0.0001
		Period 2	21%	16%	4%	< 0.0001
		Period 3	24%	8%	16%	< 0.0001

The before COVID period is January 1, 2016–December 31, 2019. The during COVID period is March 1, 2020–July 31, 2021. Arrest columns show proportion of cohort with felony or violent arrest during time period.

Table 8-2 compares the performance of the models before and during COVID-19. The AUCs are shown. Although there is some degradation in model accuracy, few of the differences are significant at $p < 0.001$. Additionally, as noted above, given differences in the length of observation periods for the before COVID-19 and post-COVID samples, results for the Period 2 and Period 3 models limit the ability to draw inferences from the findings.

Table 8-2. Model Performance Comparison for the Before and During COVID Periods

Supervision Type	Sex	Model	AUC		Difference	P-Value
			Before COVID	During COVID		
Parole	Women	Period 1	0.872	0.617	0.255	0.229
		Period 2	0.874	0.638	0.236	< 0.0001
		Period 3	0.906	0.822	0.084	0.04
	Men	Period 1	0.839	0.702	0.137	0.018
		Period 2	0.855	0.712	0.143	< 0.0001
		Period 3	0.889	0.784	0.105	< 0.0001
Split probation	Women	Period 1	0.835	0.706	0.129	0.114
		Period 2	0.8	0.782	0.018	0.591
		Period 3	0.847	0.841	0.006	0.819
	Men	Period 1	0.812	0.725	0.087	0.011
		Period 2	0.743	0.745	-0.002	0.896
		Period 3	0.797	0.738	0.059	< 0.0001
Straight probation	Women	Period 1	0.861	0.884	-0.023	0.578
		Period 2	0.81	0.829	-0.019	0.06
		Period 3	0.853	0.809	0.044	0.004
	Men	Period 1	0.83	0.792	0.038	0.239
		Period 2	0.801	0.795	0.006	0.282
		Period 3	0.822	0.793	0.029	0.003

The before COVID period is January 1, 2016–December 31, 2019. The during COVID period is March 1, 2020–July 31, 2021.

8.2 Post-COVID Validation

Table 8-3 shows the prevalence rates of felony and violent misdemeanor arrest for the pre-COVID period compared to the post-COVID period by supervision type, sex, and model period. For Period 1, the differences in arrest rates between the two periods are minimal, ranging from -2% (more arrests in the post-COVID period) to 1% (fewer arrests in the post-COVID period). As expected, given the differences in observation periods, more substantial differences occur in the Period 2 and 3 model periods.

Table 8-3. Prevalence of Felony or Violent Arrest for the Before and Post-COVID Periods

Supervision Type	Sex	Period	Felony or Violent Misdemeanor Arrest		Difference	P-Value
			Before COVID	Post-COVID		
Parole	Women	Period 1	7%	6%	1%	0.606
		Period 2	16%	9%	7%	0.001
		Period 3	13%	3%	11%	0.022
	Men	Period 1	9%	7%	2%	0.008
		Period 2	22%	13%	9%	< 0.0001
		Period 3	23%	3%	20%	< 0.0001
Split probation	Women	Period 1	10%	10%	0%	1
		Period 2	21%	12%	9%	< 0.0001
		Period 3	25%	3%	22%	< 0.0001
	Men	Period 1	9%	10%	-1%	0.019
		Period 2	24%	17%	7%	< 0.0001
		Period 3	32%	4%	28%	< 0.0001
Straight probation	Women	Period 1	12%	13%	-1%	0.272
		Period 2	18%	12%	7%	< 0.0001
		Period 3	19%	2%	16%	< 0.0001
	Men	Period 1	13%	15%	-2%	< 0.0001
		Period 2	21%	14%	7%	< 0.0001
		Period 3	24%	3%	21%	< 0.0001

The before COVID period is January 1, 2016–December 31, 2019. The post-COVID period is July 1, 2021–September 30, 2022.

Table 8-4 provides the AUCs and compares the differences in the AUCs for the before and post-COVID periods. As shown, the differences are minor, and in the one instance (Period 2 for men on parole) where the p-value is less than 0.001, the model actually performs significantly better. Overall, the results suggest that the models developed using pre-COVID data are still accurate in predicting rearrest or revocation in a post-COVID setting.

Table 8-4. Model Performance Comparison for the Before and Post-COVID Periods

Supervision Type	Sex	Period	AUC		P-Value	Difference
			Before COVID	Post-COVID		
Parole	Women	Period 1	0.830	0.811	0.695	-0.019
		Period 2	0.893	0.944	0.007	0.051
		Period 3	0.871	0.952	0.156	0.081
	Men	Period 1	0.810	0.774	0.038	-0.036
		Period 2	0.871	0.942	0.000	0.071
		Period 3	0.833	0.773	0.413	-0.060
Split probation	Women	Period 1	0.797	0.792	0.852	-0.005
		Period 2	0.844	0.875	0.170	0.031
		Period 3	0.834	0.923	0.013	0.089
	Men	Period 1	0.740	0.726	0.286	-0.014
		Period 2	0.794	0.824	0.004	0.030
		Period 3	0.810	0.775	0.422	-0.035
Straight probation	Women	Period 1	0.807	0.803	0.713	-0.004
		Period 2	0.851	0.844	0.585	-0.007
		Period 3	0.860	0.793	0.258	-0.067
	Men	Period 1	0.799	0.804	0.485	0.005
		Period 2	0.820	0.839	0.004	0.019
		Period 3	0.830	0.778	0.173	-0.052

8.3 Model Validation Summary

Model validation is the process of re-testing the predictive accuracy of algorithms on new data to ensure that the model performance has not deteriorated over time or due to changes in the underlying data. The predictive modeling process has several facets that can change over time and alter the usefulness of a predictive algorithm. First, the underlying data can change, revealed when differences are substantive in the outcome measures (here, felony and violent rearrest or revocation) or the variables used in prediction. Furthermore, the relationship between the outcomes and predictor variables

(e.g., the association between employment and rearrest) could change over time and require refitting models to reflect the current circumstances.

The COVID-19 pandemic presented a unique opportunity to assess how models built before the pandemic performed both during this period and after widespread availability of COVID-19 vaccines. However, these time periods artificially truncate the available observation time, making the first 90 days the most comparable period across the pre-, during, and post-COVID periods. Surprisingly, rearrest during the first 90 days for new starts across the periods was similar. While models developed before COVID-19 did not perform as well during COVID, the pre-COVID models were similarly accurate with the post-COVID data.

9. Limitations, Conclusions, and Recommendations

This ambitious project was designed to develop new risk assessment algorithms for the Georgia DCS that would improve the accuracy of currently used models. The goal was to develop a set of algorithms for men and women on probation, split probation, and parole that predicted the likelihood of a felony or violent misdemeanor arrest or revocation. The project was a collaborative research partnership with DCS that involved frequent interaction and consultation, culminating in the integration of the new models into the DCS CMS with a goal of deploying the new models in 2025.

Integral to the project was the goal to determine whether emerging new ML methods would provide substantive improvements over traditional statistical models such as logistic regression. ML models have been widely touted for providing enhanced prediction accuracy, but have also been criticized for a lack of transparency with respect to the underlying factors and relationships embedded in the predictions. After testing a variety of classification and survival ML models, the short answer was that although the models sometimes offered modest improvements, these improvements were insufficient to replace traditional logistic regression models, where the factors driving the outcomes were transparent, well understood, and easy to adapt into an operational setting.

A second goal—and one responsive to requests from DCS officers—was to improve the ability of models to reflect changes in risk over time and, importantly, to reflect when an individual's trajectory and risk profile was improving. Survival models provide a statistical approach to looking at changes in risk over time and allow for the inclusion of dynamic variables that take on different values over time. After investigating survival modeling approaches, these were also discarded for a multi-period set of logistic

regressions that capture three risk periods aligned with DCS policies and practices: an initial first quarter of supervision, the next three quarters of supervision in the first year, and a post-one-year period of supervision. The first quarter (Period 1) aligns with intake to supervision, the next three quarters with an initial period of standard supervision (which may be elevated or reduced), and the post-one-year period often means a reduction in supervision level to contact status. These three models provide a dynamic feature to the risk assessment with changes in risk level as individuals progress on supervision, even if there are little or no changes in the factors included in the model.

Additional dynamic capacity is reflected in the inclusion of dynamic variables that change over time. These factors can generate increased probability of the outcome (e.g., a new positive drug test) or reduced probability of the outcome (e.g., if the individual becomes employed). Identifying factors that were associated with reduced risk was somewhat hampered by the nature of supervision data collection—i.e., the tendency of agencies to record negative behavior more regularly than positive behavior. Thus, for example, employment could not be included in the Period 3 models because of the unreliability of the employment indicators in the DCS data later in supervision.

The final models perform well on standard metrics assessing prediction accuracy (i.e., AUCs in the good and usually excellent range). Further, supplemental analyses suggest that these models that were developed using data before the COVID-19 pandemic perform well on post-COVID data, providing assurance that they are valid to implement as DCS is currently working to do.

9.1 Limitations

As with any study using criminal justice agency data, issues occur with missingness and data error. Considerable effort (and discussion with our agency partners) was expended to resolve these issues while minimizing the need to exclude data from the analyses.

One notable limitation that was driven by the constraints of the data was the decision to use arrest records as opposed to conviction records in the time-specific criminal history measures. Arrests do not necessarily equate to criminal behavior, as charges may be dropped and cases may lead to *not guilty* verdicts. Further, using arrests as an indicator of prior criminality may disproportionately impact communities of color, as bias within policing practices can influence arrest rates. As noted in the data section, this decision was made for several reasons;

1. Conviction data is provided by the Georgia Crime Information Center (GCIC) and conviction data traditionally lags behind arrest records and is updated semi-annually by the GCIC. Although this would have minimal impact on convictions dating back more than a year, the underlying charges and convictions for the supervision term would be incomplete and risk scores would potentially be impacted throughout the course of supervision based on updated conviction data as opposed to real changes in the nature of someone's progress on supervision.
2. Completeness of conviction data from the GCIC varies throughout the state of Georgia. Internal reporting provided by Georgia DCS indicated that the completeness of charge disposition data associated with arrest records varied substantially by the county in which the charges were being tried. This lack of

complete disposition data could lead to biased results depending on reporting rates of the counties in which the charges were prosecuted.

Given the issues associated with using conviction data, the study team decided to use the more complete arrest data. To alleviate some of the issues associated with racial bias in arrest records, these models used race as a predictor in the training data models but omitted race from actual predictions. This technique, while imperfect, accounts for some of the variance in arrest records attributed to different base arrest rates by racial group without influencing predictions. In addition, it is possible that while models built with arrest records performed well, those built with conviction data would have performed equally well or better. However, given the limitations for conviction data, this study uses available arrest data.

Similarly, this model accounts for racial differences primarily by examining the differences in predictions between individuals who are white vs. non-white. One limitation is that this does not incorporate detail on racial and ethnic differences within the individuals aggregated to the non-white category. However, across supervision types roughly 93%-95% of the non-white categorization was comprised of Black or African American individuals, which limits the ability to draw inference for non-white individuals of other race or ethnicities. As such, further research exploring these distinctions is needed in research sites where the non-white supervised population has greater diversity and sample sizes will allow for meaningful sub-group analyses.

It should also be noted that, while the AUC measures for all models were high, there were fluctuations in accuracy by race. This is likely due to small sample sizes for females in certain supervision types, specifically in the test datasets for parole, where

rearrest was a rare outcome in certain time periods. Future research could allow for a longer observation period for new supervision starts that may bolster sex- and supervision type-specific models by increasing sample size.

Another limitation was the relative lack of data that could be used to predict improvements in risk. If agencies are interested in predicting positive outcomes for those on supervision, they are encouraged to begin more routinely collecting information on those prosocial factors theorized to be associated with desistance—such as stable employment or housing—so that these measures can be included in future model development.

An additional limitation is the use of AUC to measure model accuracy. While this statistic is a robust threshold neutral metric to assess model performance, it provides the total model accuracy, as opposed to assessing model performance below or above specific thresholds, commonly referred to as partial AUC metrics. pAUC tests are common when assessing model performance when producing high or low predicted probabilities. Future analyses should incorporate partial accuracy measures as part of a robustness-testing effort.

Further, as with any risk model, these models are predictions that have error. To our knowledge, we are the first research team to estimate error bands around risk predictions for a criminal justice population. Although not exhaustively described herein, subsequent publications are intended to initiate a conversation among fellow researchers and practitioners about the fact that the error associated with each factor contributes to the overall error, implying that the accuracy of a prediction of, for

example, 0.3, may vary substantially depending upon the source (factors) predicting that 0.3.

The development of these models yielded many important and practical findings for practitioners. Specifically, in this setting using arrests in lieu of convictions was effective, although this was largely driven by the limitations and timeliness of the statewide conviction data. Additionally, the IDRACS models showed that limiting the recall period for arrests to the previous five years did not worsen model performance. Further, although using select machine learning classifiers did yield limited improvements in some time periods, the slight increase in accuracy did not outweigh the loss of interpretability and ease of implementation, which is a key feature of traditional logistic regression classification models. While the IDRACS models were tested extensively using data from different time periods in Georgia, it should be noted that these models reflect supervision and data collection practices in the state of Georgia. For researchers or practitioners wishing to develop similar models, sufficient testing should be done within the research site to ensure that the model fits both local data and supervision practices.

References

Agresti, A. (2012). *Categorical Data Analysis* (Vol. 792). John Wiley & Sons.

Altman, D. G., & Bland, J. M. (1998). Time to event (survival) data. *BMJ*, *317*(7156), 468–469. https://doi.org/10.1136/BMJ.317.7156.468

Andrews, D. A., Bonta, J., & Wormith, J. S. (2000). *Level of Service/Case Management Inventory*. Multi-Health Systems. https://doi.org/10.1037/t05029-000

Andrews, D. A., Bonta, J., & Wormith, J. S. (2011). The Risk-Need-Responsivity (RNR) Model: Does adding the good lives model contribute to effective crime prevention? *Criminal Justice and Behavior*, *38*(7), 735–755. https://doi.org/10.1177/0093854811406356

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, *23*, 77–91.

Baglivio, M. (2007). *The prediction of risk to recidivate among a juvenile offending population* [University of Florida]. https://www.assessments.com/assessments_documentation/PACT%20Validation%20Dissertation%20Summary%20by%20Dr.%20Michael%20Baglivio.pdf

Baglivio, M., & Jackowski, K. (2012). Examining the Validity of a Juvenile Offending Risk Assessment Instrument Across Gender and Race/Ethnicity. *Youth Violence and Juvenile Justice*, *11*(1), 26–43. https://doi.org/10.1177/1541204012440107

Baldwin, R. M., Owzar, K., Zembutsu, H., Chhibber, A., Kubo, M., Jiang, C., Watson, D., Eclov, R. J., Mefford, J., McLeod, H. L., Friedman, P. N., Hudis, C. A., Winer, E. P., Jorgenson, E. M., Witte, J. S., Shulman, L. N., Nakamura, Y., Ratain, M. J., & Kroetz, D. L. (2010a). R: A language and environment for statistical computing. *(No Title)*, *18*(18), 5099–5109. https://doi.org/10.1158/1078-0432.CCR-12-1590

Baldwin, R. M., Owzar, K., Zembutsu, H., Chhibber, A., Kubo, M., Jiang, C., Watson, D., Eclov, R. J., Mefford, J., McLeod, H. L., Friedman, P. N., Hudis, C. A., Winer, E. P., Jorgenson, E. M., Witte, J. S., Shulman, L. N., Nakamura, Y., Ratain, M. J., & Kroetz, D. L. (2010b). R: A language and environment for statistical computing. *(No Title)*, *18*(18), 5099–5109. https://doi.org/10.1158/1078-0432.CCR-12-1590

Baumgartner, P., Godwin, A., & Hadley, E. (2021). *Rapid Offense Text Autocoder | RTI*. https://www.rti.org/insights/rapid-offense-text-autocoder

Berk, R. A., & Bleich, J. (2013). Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment. *Criminology & Public Policy*, *12*. https://heinonline.org/HOL/Page?handle=hein.journals/crpp12&id=529&div=&collection=

Berk, R. A., Kriegler, B., & Baek, J. H. (2006). Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology*, *22*(2), 131–145. https://doi.org/10.1007/S10940-006-9005-Z/METRICS

Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions. *Journal of Empirical Legal Studies*, *13*(1), 94–115. https://doi.org/10.1111/JELS.12098

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Https://Doi.Org/10.1177/0049124118782533*, *50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). *VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT Prepared for the*. Florida State University.

Boman, J. H., & Gallupe, O. (2020). Has COVID-19 Changed Crime? Crime Rates in the United States during the Pandemic. *American Journal of Criminal Justice*, *45*(4), 537–545. https://doi.org/10.1007/S12103-020-09551-3/METRICS

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harlend (Ed.), *Choosing Correctional Options that Work: Defining the Demand and Evaluating the Supply.* . Sage. https://psycnet.apa.org/record/1996-97573-001

Bonta, J., & Andrews, D. A. (2007). Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation 2007-06. *Rehabilitation*, *6*(1), 1–22. https://www.securitepublique.gc.ca/cnt/rsrcs/pblctns/rsk-nd-rspnsvty/rsk-nd-rspnsvty-eng.pdf

Bonta, J., & Wormith, S. J. (2007). Risk and needs assessment. In G. McIvor & P. Raynor (Eds.), *Developments in Social Work Offenders* (Vol. 48, pp. 131–152).

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). A review of survival trees. *Https://Doi.Org/10.1214/09-SS047*, *5*(none), 44–71. https://doi.org/10.1214/09-SS047

Brame, R., Bushway, S. D., & Paternoster, R. (2003). EXAMINING THE PREVALENCE OF CRIMINAL DESISTANCE*. *Criminology*, *41*(2), 423–448. https://doi.org/10.1111/J.1745-9125.2003.TB00993.X

Brennan, T., Dieterich, W., & Ehret, B. (2008). 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

Brouillette-Alarie, S., & Proulx, J. (2013). Predictive validity of the Static-99R and its dimensions. *Journal of Sexual Aggression*, *19*(3), 311–328. https://doi.org/10.1080/13552600.2012.747630

Brown, S. L., St. Amand, M. D., & Zamble, E. (2009). The dynamic prediction of criminal recidivism: A three-wave prospective study. *Law and Human Behavior*, *33*(1), 25–45. https://doi.org/10.1007/S10979-008-9139-7/METRICS

Bureau of Justice Assistance. (n.d.). *History of Risk Assessment*.

Burrell, W. D. (2016). Risk and needs assessment in probation and parole. In F. S. Taxman (Ed.), *Handbook on Risk and Need Assessment: Theory and Practice* (pp. 39–64). Routledge. https://doi.org/10.4324/9781315682327

Campbell, C., Papp, J., Barnes, A., Onifade, E., & Anderson, V. (2018). Risk Assessment and Juvenile Justice: An interaction between risk, race, and gender. *Criminology & Public Policy*, *17*(3), 525–545. https://doi.org/10.1111/1745-9133.12377

Cardoso, R. L., Almeida, V., Meira, W., & Zaki, M. J. (2019). A framework for benchmarking discrimination-aware models in machine learning. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 437–444. https://doi.org/10.1145/3306618.3314262

Carson, E. A. (2021). *Prisoners in 2021 - Statistical Tables*. https://www.ojp.gov/library/publications/prisoners-2021-statistical-tables

Caudy, M. S., Durso, J. M., & Taxman, F. S. (2013). How well do dynamic needs predict recidivism? Implications for risk assessment and risk reduction. *Journal of Criminal Justice*, *41*(6), 458–466. https://doi.org/10.1016/J.JCRIMJUS.2013.08.004

Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and Mathematical Methods in Medicine*, *2013*. https://doi.org/10.1155/2013/873595

Chung, C. F., Schmidt, P., & Witte, A. D. (1991). Survival analysis: A survey. *Journal of Quantitative Criminology*, *7*(1), 59–98. https://doi.org/10.1007/BF01083132/METRICS

Cohen, T. H., & Lowenkamp, C. T. (2018). Revalidation of the Federal PTRA: Testing the PTRA for Predictive Biases. *Https://Doi.Org/10.1177/0093854818810315*, *46*(2), 234–260. https://doi.org/10.1177/0093854818810315

Connolly, M. (2003). *A Critical Examination of Actuarial Offender-Based Prediction Assessments: Guidance for the Next Generation of Assessments* . The University of Texas at Austin.

Craig, E., Zhong, C., & Tibshirani, R. (2021). *Survival stacking: casting survival analysis as a classification problem*. https://arxiv.org/abs/2107.13480v1

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., & Comfort, M. (2020). Public safety assessment. *Criminology & Public Policy*, *19*(2), 409–431. https://doi.org/10.1111/1745-9133.12481

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2017). Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. *Handbook of Recidivism Risk/Needs Assessment Tools*, 1–29. https://doi.org/10.1002/9781119184256.CH1

Dyck, H., Campbell, M., & Wershler, J. (2018). Real-world use of the risk–need–responsivity model and the level of service/case management inventory with community-supervised offenders. *Law and Human Behavior*, *42*(3), 258. https://psycnet.apa.org/record/2018-14329-001

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage.

Fratello, J., Salsich, A., & Mogulescu, S. (2011). Juvenile Detention Reform in New York City: Measuring Risk Through Research. *Federal Sentencing Reporter*, *24*(1), 15–20. https://doi.org/10.1525/FSR.2011.24.1.15

Giunchiglia, E., Nemchenko, A., & van der Schaar, M. (2018). RNN-SURV: A deep recurrent model for survival analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11141 LNCS*, 23–32. https://doi.org/10.1007/978-3-030-01424-7_3/COVER

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, *2017-January*, 79–85. https://doi.org/10.1109/ICACCI.2017.8125820

Greiner, L. E., Law, M. A., & Brown, S. L. (2014). Using Dynamic Factors to Predict Recidivism Among Women. *Http://Dx.Doi.Org/10.1177/0093854814553222*, *42*(5), 457–480. https://doi.org/10.1177/0093854814553222

Hanson, R. K., Thornton, D., Helmus, L. M., & Babchishin, K. M. (2016). What Sexual Recidivism Rates Are Associated With Static-99R and Static-2002R Scores? *Sexual Abuse*, *28*(3), 218–252. https://doi.org/10.1177/1079063215574710

Harcourt, B. E. (2015). Risk as a Proxy for Race. *Federal Sentencing Reporter*, *27*(4), 237–243. https://doi.org/10.1525/FSR.2015.27.4.237

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-21606-5

Hastie, T., & Stanford, J. Q. (2016). *Glmnet Vignette*. http://cran.us.r-project.org

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute Recidivism Rates Predicted By Static-99R and Static-2002R Sex Offender Risk Assessment Tools Vary Across Samples. *Criminal Justice and Behavior*, *39*(9), 1148–1171. https://doi.org/10.1177/0093854812443648

Helmus, L. M., Kelley, S. M., Frazier, A., Fernandez, Y. M., Lee, S. C., Rettenberger, M., & Boccaccini, M. T. (2022). Static-99R: Strengths, Limitations, Predictive Accuracy Meta-Analysis, and Legal Admissibility Review. *Psychology, Public Policy, and Law*. https://doi.org/10.1037/LAW0000351

Hodgkinson, T., & Andresen, M. A. (2020). Show me a man or a woman alone and I'll show you a saint: Changes in the frequency of criminal incidents during the COVID-19 pandemic. *Journal of Criminal Justice*, *69*, 101706. https://doi.org/10.1016/J.JCRIMJUS.2020.101706

Homepage | Advancing Pretrial Policy & Research (APPR). (n.d.). Retrieved December 20, 2023, from https://advancingpretrial.org/

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Https://Doi.Org/10.1214/08-AOAS169*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Jahn, J. L., Simes, J. T., Cowger, T. L., & Davis, B. A. (2022). Racial Disparities in Neighborhood Arrest Rates during the COVID-19 Pandemic. *Journal of Urban Health*, *99*(1), 67–76. https://doi.org/10.1007/S11524-021-00598-Z/FIGURES/3

Jimenez, A. C., Delgado, R. H., Vardsveen, T. C., & Wiener, R. L. (2018). Validation and Application of the LS/CMI in Nebraska Probation. *Criminal Justice and Behavior*, *45*(6), 863–884. https://doi.org/10.1177/0093854818763231

Kaeble, D. (2021). *Probation and Parole in the United States, 2020*.

Kaeble, D. (2023). *Probation and Parole in the United States, 2021*. https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/ppus21.pdf

Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T., & Kluger, Y. (2016). DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, *18*(1). https://doi.org/10.1186/s12874-018-0482-1

Kleinbaum, D. G., & Klein, M. (2012). *The Cox Proportional Hazards Model and Its Characteristics*. 97–159. https://doi.org/10.1007/978-1-4419-6646-9_3

Kremers, W. K. (2007). *Concordance for Survival Time Data: Fixed and Time-Dependent Covariates and Possible Ties in Predictor and Time*.

Kurlychek, M. C., Brame, R., & Bushway, S. D. (2006). SCARLET LETTERS AND RECIDIVISM: DOES AN OLD CRIMINAL RECORD PREDICT FUTURE OFFENDING?*. *Criminology & Public Policy*, *5*(3), 483–504. https://doi.org/10.1111/J.1745-9133.2006.00397.X

Latessa, E. J., Lemke, R., Makarios, M., & Smith, P. (2010). The Creation and Validation of the Ohio Risk Assessment System (ORAS). *Federal Probation*, *74*. https://heinonline.org/HOL/Page?handle=hein.journals/fedpro74&id=16&div=&collection=

Latessa, E. J., & Lovins, B. (2010). The Role of Offender Risk Assessment: A Policy Maker Guide. *Victims and Offenders*, *5*(3), 203–219. https://doi.org/10.1080/15564886.2010.485900

Lovins, B. K., Latessa, E. J., May, T., & Lux, J. (2018). Validating the Ohio Risk Assessment System Community Supervision Tool with a Diverse Sample from Texas. *Corrections*, *3*(3), 186–202. https://doi.org/10.1080/23774657.2017.1361798

McCafferty, J. T. (2016). The importance of counties: Examining the predictive validity of a state juvenile risk assessment instrument. *Journal of Offender Rehabilitation*, *55*(6), 377–395. https://doi.org/10.1080/10509674.2016.1194944

Milgram, A., Holsinger, A. M., Vannostrand, M., & Alsdorf, M. W. (2014). Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making. *Federal Sentencing Reporter*, *27*. https://heinonline.org/HOL/Page?handle=hein.journals/fedsen27&id=224&div=&collection=

Mueller, K. C., Carey, M. T., & Noh, K. (2022). Revalidating the Positive Achievement Change Tool (PACT) in a Major City: A Survival Analysis. *Crime & Delinquency*, *69*(13–14), 2678–2698. https://doi.org/10.1177/00111287221087949

Muthukrishnan, R., & Rohini, R. (2017). LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*, 18–20. https://doi.org/10.1109/ICACA.2016.7887916

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC), 63623. https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX

Ogloff, J. R. P., & Davis, M. R. (2004). Advances in offender assessment and rehabilitation: Contributions of the risk–needs–responsivity approach. *Psychology, Crime & Law*, *10*(3), 229–242. https://doi.org/10.1080/10683160410001662735

Olver, M. E., Beggs Christofferson, S. M., Grace, R. C., & Wong, S. C. P. (2013). Incorporating Change Information Into Sexual Offender Risk Assessments Using the Violence Risk Scale–Sexual Offender Version. *Sexual Abuse*, *26*(5), 472–499. https://doi.org/10.1177/1079063213502679

Onifade, E., Davidson, W., & Campbell, C. (2009). Risk Assessment: The Predictive Validity of the Youth Level of Service Case Management Inventory with African Americans and Girls. *Journal of Ethnicity in Criminal Justice*, *7*(3), 205–221. https://doi.org/10.1080/15377930903143544

Park, S. Y., Park, J. E., Kim, H., & Park, S. H. (2021). Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches). *Korean Journal of Radiology*, *22*(10), 1697. https://doi.org/10.3348/KJR.2021.0223

Penciana, M. J., & D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, *23*(13), 2109–2123. https://doi.org/10.1002/SIM.1802

Perkins, C. (1993). *National Corrections Reporting Program*.

Phenix, A., Helmus, L., & Hanson, R. K. (2016). *Static-99R & Static-2002R evaluators' workbook*.

Ranganath, R., Perotte, A., Elhadad, N., & Blei, D. (2016). *Deep Survival Analysis* (pp. 101–114). PMLR. https://proceedings.mlr.press/v56/Ranganath16.html

Raynor, P. (2016). Three narratives of risk: corrections, critique and context. *Beyond the Risk Paradigm in Criminal Justice*, 24–46. https://link.springer.com/content/pdf/10.1057/978-1-137-44133-1.pdf#page=33

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 1–8. https://doi.org/10.1186/1471-2105-12-77/TABLES/3

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review. *Behavioral Sciences & the Law*, *31*(1), 55–73. https://doi.org/10.1002/BSL.2053

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

Skeem, J.L., Monahan, J., & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior, 40*(5), 580–593. https://doi.org/10.1037/lhb0000206

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/J.ESWA.2019.05.028

Steadman, H. J., Robbins, P. C., Islam, T., & Osher, F. C. (2007). Revalidating the brief jail mental health screen to increase accuracy for women. *Psychiatric Services*, *58*(12), 1598–1601. https://doi.org/10.1176/PS.2007.58.12.1598/ASSET/IMAGES/LARGE/JC17T1.JPEG

Travis, J., Western, B., & Redburn, F. (2014). The Growth of Incarceration in the United States: Exploring Causes and Consequences. In *Publications and Research*. National Research Council. https://academicworks.cuny.edu/jj_pubs/27

Vincent, G. M., & Viljoen, J. L. (2020). Racist Algorithms or Systemic Problems? Risk Assessments and Racial Disparities. *Criminal Justice and Behavior*, *47*(12), 1576–1584. https://doi.org/10.1177/0093854820954501

Winokur-Early, K., Hand, G. A., & Blankenship, J. L. (2012). *Validity and reliability of the Florida Positive Achievement Change Tool (PACT) risk and needs assessment instrument: A three-phase evaluation (validation study, factor analysis, inter-rater reliability)*

Xu, J., & Long, J. S. (2005). Using the Delta Method to Construct Confidence Intervals for Predicted Probabilities, Rates, and Discrete Changes. *The Stata Journal*, *5*(4), 537–559. www.indiana.edu/~jslsoc/spost.htm

Yukhnenko, D., Blackwood, N., & Fazel, S. (2020). Risk factors for recidivism in individuals receiving community sentences: a systematic review and meta-analysis. *CNS Spectrums*, *25*(2), 252–263. https://doi.org/10.1017/S1092852919001056

Zeng, Z. (2022). *Jail Inmates in 2021 - Statistical Tables*. https://www.ojp.gov/library/publications/jail-inmates-2020-statistical-tables

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. M. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine*, *6*(7), 121–121. https://doi.org/10.21037/ATM.2018.02.12

Appendix A

Table A-1. Terms Used to Detect Probation Conditions in Court Docket “Special Conditions” Field

Condition Type	Search Terms
Employment	"employment", "job"
Education	"ged", "high school diploma", "vocational training"
Fee	"fee", "fees", "fine", "restitution", "pay", "paid", "financial", "payment"
No contact	"no contact", "stay away from", "shall not enter", "banished", "banned", "avoid contact"
Drug or alcohol (general)	“drug”, “alcohol”, “s a eval”, “sa eval”, “submit specimen”, “specimen”, “dna sample”, “a d eval”, “d a eval”, “ad eval”, “da eval”, “alcohol testing”, “alcohol test”, “drug testing”, “drug test”, “substance abuse”, “substance abuse eval”, “substance eval”, “substance abuse evaluation”, “alcohol drug”, “drug conditions”, “a d conditions”, “ad conditions”
Drug or alcohol (treatment)	“rsat”, “substance abuse treatment”, “residential substance abuse”, “drug treatment”, “treatment”, “drc”, “day reporting center”, “eval and treatment”, “evaluation and treatment”, “eval treatment”, “evaluation treatment”, “rehab”
Sex offender	“sex offender”
Community service	“community service”, “csw”, “comm service”, “cs”
Violence-related	“violent”, “violence”, “harass”, “threaten”, “intimidate”, “weapons”, “weapon”, “firearm”, “firearms”, “anger mgmt”, “anger management”
Other	“curfew”, “no driving”, “4th amendment”, “fourth amendment”, “4th waiver”, “4th amend”, “ankle monitor”, “em”

Appendix B

The Decay Effect of Criminal History

In addition to identifying predictive models that fit specific time periods, we also sought to produce parsimonious models, as simpler models were less likely to overfit new data. As part of this process, we examined the utility of using complete criminal history measures, as opposed to measures that limited the recall period, without compromising predictive accuracy. This approach sought to identify how far back one needs to look at prior criminal arrests to assist in predicting future arrests. The effect on prediction was shown in Table 6-1. To achieve this, we employed feature selection algorithms and statistical tests of differences in predictive accuracy, as described below.

LASSO and AUC Results

LASSO regression can be used in factor selection to identify factors that are not explaining substantial variance. This information can be used to reduce the number of factors included in a model without compromising accuracy, thus making the model more parsimonious and likely able to fit new data better (Muthukrishnan & Rohini, 2017). To assist in identifying measures that could be dropped, we employed LASSO logistic regression on models that included all dynamic features and all time-specific criminal history measures. These criminal history measures were counts of arrest charges for violent, property, drug, public order, and probation/parole offenses, as well as prison terms that occurred 0–1 year, 1–2 years, 2–5 years, 5–10 years, and 10+ years before the start of supervision. This full set of criminal history measures includes 30 variables or factors. The goal was to determine whether this number could be reduced—increasing parsimony—without compromising predictive accuracy.

LASSO regression models were estimated for Period 1 for men on straight probation (N = 24,070) and split probation (N = 10,724). For the Period 1 straight probation model including all 30 criminal history measures, the AUC was 0.776. LASSO results suggested that details on drug or violent offenses or prior prison terms beyond 5 years before the start of supervision could be omitted from the models without compromising predictive accuracy. The AUC for the LASSO model including 12 factors was 0.78. Similarly, for men on split probation, the LASSO model penalized out drug, violent, and property charges that were between 5–10 years old and drug, public order, and violent offenses that were more than 10 years old. For this model with all criminal history factors, the AUC was 0.747, which is not significantly different from the LASSO 12-factor model that had an AUC of 0.735. Thus, these models suggested that the lookback period could be reduced while retaining the same predictive accuracy.

Based on the results of LASSO models applied to different time periods by sex and supervision type, we specified our models using a 5-year lookback period (from the start of supervision) for the first year of supervision (Periods 1 and 2) and a 2-year lookback period after the first year (Period 3). Table B-1 shows the results of statistical tests that compare two paired AUC values. In most cases, the more parsimonious model produced slightly more accurate results or no statistical difference in prediction. This suggests that, in this sample, there is limited predictive value in looking past 5 years when including criminal history.

Table B-2. Model AUC Comparisons Using Limited Lookback vs. All-Time Criminal History

Supervision Type	Sex	Period 1			Period 2			Period 3		
		Limited Lookback	All Time		Limited Lookback	All Time		Limited Lookback	All Time	
		AUC	AUC	P-Value	AUC	AUC	P-Value	AUC	AUC	P-Value
Straight probation	Men	0.801	0.798	0.011	0.825	0.822	0.021	0.828	0.829	0.673
	Women	0.801	0.798	0.147	0.851	0.850	0.489	0.855	0.858	0.207
Split probation	Men	0.748	0.748	0.863	0.795	0.796	0.639	0.789	0.802	0.000
	Women	0.759	0.750	0.307	0.816	0.820	0.537	0.833	0.826	0.354
Parole	Men	0.812	0.807	0.392	0.858	0.855	0.120	0.819	0.819	0.933
	Women	0.731	0.717	0.309	0.901	0.890	0.263	0.813	0.830	0.514