PPIC | PUBLIC POLICY INSTITUTE OF CALIFORNIA

25 YEARS

Pretrial Risk Assessment in California

Technical Appendices

CONTENTS

Heather M. Harris, Justin Goss, and Alexandria Gumbs

Appendix A. Case Studies

Introduction

In this supplementary section, we present six two-page case studies that draw on the work of local governments, journalists, and researchers who encountered and overcame challenges as they sought to implement or understand pretrial risk assessments tools. For each case, we provide a summary and several key "takeaways." The cases provide concrete examples that support the key points made in main report.

In the first case we describe how officials in Riverside County updated their Pretrial Services Division. The case illustrates the advantages of locally validating and modifying a pre-existing risk assessment tool. It also highlights the need for jurisdictions to clearly define the objectives they want to achieve through their pretrial risk assessment systems because technical and policy decisions made during the development of those systems can either promote or undercut those objectives (CJI 2017; Lovins and Lovins 2015).

The Santa Clara County case allows us to highlight several key points from the main report. Risk level classifications from pretrial risk assessment tools can be misleading, which highlights the importance of developing policy frameworks—what we call pretrial risk assessment systems—to transparently and consistently translate risk level classifications into pretrial release or detention recommendations. The county's commitment to routine monitoring and regular evaluation of its pretrial risk assessment system also highlights the importance of addressing overrides. Overrides can negatively impact the transparency, consistency, and equity of pretrial release or detention decisions. To understand why overrides occur, what their impact is, and how they can be addressed, the reasons for overrides must be consistently recorded (BRWG 2016; Levin 2012).

We then describe how Sonoma County created a pretrial risk assessment tool and system, highlighting the advantages of a transparent local process. We discuss the challenges of such an ambitious undertaking, including how to define and measure risk, whether to include predictors such as socioeconomic and mental health status, and what can happen when pretrial risk assessment tools classify too many people as medium risk. In addition, Sonoma County's recent evaluation of its pretrial risk assessment tool and system provide an excellent example for other counties to follow (PJI n.d.; Feld and Halverson 2019; Robertson and Jones 2013).

Next we describe the work of two researchers who developed their own pretrial risk assessment tool using only two predictors. Their work demonstrates that even jurisdictions with limited resources or data, may be able to develop a pretrial risk assessment tool for use within a pretrial risk assessment system (Dressel and Farid 2018).

Our final two case studies focus on equity—and whether it can be achieved by any measure—that have been raised as the use of pretrial risk assessment tools has proliferated. We begin by describing ProPublica's conflict with Northpointe, the proprietors of the COMPAS. Their disagreement illustrates, first, that there are multiple ways to quantitatively define equity; second, that not all definitions of equity can be satisfied simultaneously; and third, that racial inequity originates in the historical criminal justice data that pretrial risk assessment tools rely on to make risk predictions (Angwin et al. 2016; Dieterich, Mendoza, and Brennan 2016; Mayson 2018).

Finally, we summarize findings from a recent Center for Court Innovation study. This case presents an example of how pretrial risk assessment tools and systems can be evaluated and adjusted to promote local policy objectives related to equity. It illustrates how California's counties can, first, determine the degree of racial and other forms of inequity that pretrial risk assessment tools might propagate *and,* second, how that inequity can be mitigated by developing and testing alternative policies for interpreting risk predictions and making pretrial release or detention decisions based on them (Picard et al. 2019).

Riverside Case Study

Riverside County began using the Virginia Pretrial Risk Assessment Tool (VPRAI) in 2014 and validated it locally two years later. A Pretrial Steering Committee (PSC) comprised of representatives from the Probation Department, Pretrial Services Unit, the Court, Sheriff's Department, and offices of the Public Defender and District Attorney oversaw validation.

The PSC set three clearly defined pretrial policy objectives: to release more people on own recognizance; to ensure release decisions correspond to assessed risk; and to develop a continuum of supervision options (i.e., "graduated sanctions"), from release on own recognizance for the lowest risk individuals, to detention for the highest risk individuals. To achieve these objectives the PSC validated the VPRAI locally, invested in electronic monitoring to supervise released individuals, and automated court date reminders to reduce failure to appear rates.

During validation, the VPRAI was modified to create the Riverside PRAI (RPRAI). The RPRAI maintained the same definition of pretrial misconduct—a compound outcome of either failure to appear in court or pretrial arrest—but reduced the number of predictors of pretrial misconduct from nine to five. The five predictors measured criminal history, housing status, and substance use. In addition, the number of risk level classifications was reduced from five to three. As a result of these changes the overall accuracy of the RPRAI improved slightly relative to the VPRAI, increasing from 0.609 to 0.614. However, the performance of the RPRAI varied for different demographic subgroups of individuals. The RPRAI was slightly more accurate for females than for males and for nonwhites relative to whites.

How people were classified using the RPRAI may have undermined the local policy objectives defined by the PSC because high risk individuals, on average, were still less likely to commit pretrial misconduct than not and most individuals were classified as moderate risk—common outcomes in pretrial risk assessment. Individuals classified as low risk under the RPRAI had pretrial misconduct rates of 13 percent, while those classified as moderate and high risk committed pretrial misconduct at rates of 27 percent and 43 percent respectively. Nearly 60 percent of assessed individuals fell into the moderate risk level classification, whereas 14 percent fell into the low risk level classification and 28 percent were classified as high risk. Judges overrode the pretrial release or detention recommendations from the RPRAI 30 percent of the time.

Takeaways

Validation of an existing tool can lead to performance improvements.

By adopting the VPRAI, Riverside avoided the challenges associated with developing a bespoke tool from scratch. By modifying the tool Riverside demonstrated that the local performance of VPRAI could be improved and also generated information regarding how the tool performed on different local demographic subgroups, which is crucial to assessing equity in risk prediction and pretrial release or detention decisions.

Risk level classifications should enable pretrial decisions that support policy objectives.

Only about one in ten individuals assessed using the RPRAI were classified as low risk and, thus, clearly eligible for release. This likely contributed to the county's failure to release a higher share of its pretrial population, as evidenced by rising proportions of pretrial detainees in the county jail in recent years (BSCC Jail Profile Survey). To create the conditions under which the objective of releasing more people on their own recognizance can be met, the PSC could adjust the cut points to classify more people as low risk.

Robust pretrial risk assessment systems interpret ambiguous risk level classifications.

Similarly, the RPRAI classified so many individuals as moderate risk that judges likely could not differentiate between moderate risk individuals who should be released and moderate risk individuals who should be detained. To facilitate those decisions, the PSC could provide more guidance to judges. Specifically, the conditions under which medium risk people should be released can be broadened by expanding graduated sanctioning options.

Policies should be developed to address risk assessment overrides.

The absence of a strong pretrial risk assessment system to inform pretrial release or detention decisions based on the RPRAI also likely contributed to high rates of judicial overrides. Although the county tracked overrides, it neither evaluated how those overrides impacted the accuracy and equity of the RPRAI nor responded by taking steps to minimize them. For example, the PSC could track the reasons for overrides and use that information to develop a decision matrix that relates risk level classifications to information omitted from the RPRAI. Such a framework might promote more consistency and transparency in judges' decisions.

Santa Clara County Case Study

About a decade ago, the Pretrial Justice Institute helped Santa Clara County develop a pretrial risk assessment tool that includes three risk prediction models that predict three pretrial misconduct outcomes—new arrest, failure to appear, and technical violations—for assessed individuals. A workgroup comprised of local criminal justice officials also created a pretrial risk assessment system to interpret risk predictions from the tool. The workgroup developed a scoring manual and created a decision matrix that associated risk level classifications with pretrial release or detention decisions and supervision conditions.

Santa Clara County engaged in a collaborative process to develop a pretrial risk assessment tool and a pretrial risk assessment system. Yet two aspects of the risk level classifications produced by the tool illustrate potential challenges associated with making informative classifications. First, some pretrial misconduct outcomes were rare. For example, 99 percent, 93 percent, and 89 percent of individuals classified at levels one (lowest), two, and three (highest), respectively, were not arrested during the pretrial period. As discussed in Technical Appendix C, rare outcomes are difficult to predict, which led to a second problem. Most classified individuals fell into one risk level classification—a sign that the risk prediction model could not differentiate between high and low risk individuals. For instance, 93 percent of individuals were classified at level two by the failure to appear model.

Santa Clara County evaluates its pretrial risk assessment system regularly. Those regular evaluations include examination of overrides—departures from the recommendations of the system—by judges and pretrial services officers (PSOs). According the Santa Clara County Bail and Release Workgroup, Santa Clara allows PSOs to override 15 percent of the time and only after they specify reasons for overrides, which are reviewed by a supervisor. Yet judges can override PSOs recommendations without specifying why. Judges overrode the recommendations of PSOs 25 percent of the time in 2015.[1] "Anecdotal information" indicates that judges override in response to additional information provided by the prosecutor, a process that could be formalized to account for different types of information (BRWG 2016: 45).

Takeaways

Use separate risk prediction models to predict each pretrial misconduct outcome.

According to a report from the Partnership on AI, an organization dedicated to studying best practices in artificial intelligence, different pretrial misconduct outcomes should be predicted using separate risk prediction models (PAI 2019). Yet many existing pretrial risk assessment tools predict compound outcomes (e.g., failure to appear and arrest) using a single risk prediction model. By contrast, Santa Clara County's pretrial risk assessment tool predicts three outcomes using separate risk prediction models, which allows policymakers to differentiate between risks of pretrial misconduct and to create graduated sanctions based on those differences.

Understand what "high" and "low" risk mean in the local population.

To make appropriate pretrial release or detention decisions, judges and PSOs should understand what "high" and "low" risk mean in terms of the chance that a person will commit pretrial misconduct. In Santa Clara County, pretrial misconduct was rare, which may have distorted the meaning of high risk. Only 11 percent of individuals classified as high risk were arrested after being released during the pretrial period. Put another way, people

[1] By 2019, judges' decisions were in concordance with the pretrial risk assessment system in at 90 percent of cases—although they still do not record the reasons for their overrides (personal communication 2019).

assessed at high risk had 89 percent probability of *not* being arrested. Thus, in Santa Clara County, even many individuals classified as high risk may have been safe to release.

When risk level classifications do not inform pretrial release or detention decisions, pretrial risk assessment systems should.

The pretrial risk assessment tool used in Santa Clara County classified most individuals as medium risk. In fact, the failure to appear model classified people as medium risk with such high prevalence that it provided judges with little information about how to determine who should be released and who should be detained. Although Santa Clara County developed a decision matrix to inform judges' and PSOs' release or detention decisions, those recommendations are regularly overridden—suggesting a misalignment between the risk assessment system and the individuals who make those decisions. This misalignment can be addressed by adjusting the policies within the pretrial risk assessment system to accommodate or eliminate overrides—but only if more information about them is collected.

Routinely monitor and regularly evaluate pretrial risk assessment tools and systems.

Santa Clara County routinely monitors and regularly evaluates its pretrial risk assessment system, which has resulted in higher pretrial release rates and lower pretrial misconduct rates. However, override rates have increased over time in Santa Clara County. Although the county has taken steps to address overrides, more could be done to understand why they are occurring, how they might impact consistency and equity in pretrial release or detention decisions, and to refine the pretrial risk assessment system in response.

Require judges to record why they override.

High override rates among PSOs and judges threaten the transparency, equity, and consistency of pretrial risk assessment systems. Although PSOs in Santa Clara County are required to provide their supervisor with written justifications for overrides, the same does not seem to be true for judges (BRWG 2016). Collecting data on the reasons for overrides will enable evaluators to characterize the situations in which they happen, determine whether they introduce inconsistency or inequity in the administration of pretrial justice, and redesign the pretrial risk assessment system to ameliorate or accommodate them. An example of this is Sonoma County's system of "enhancements," which is described in the following case.

Sonoma County Case Study

Sonoma County redesigned its pretrial policy framework by creating a risk assessment system around a locally developed pretrial risk assessment tool. The locus of the redesign was the Community Corrections Partnership (CCP), a local policymaking workgroup comprised of representatives from county administrative, criminal justice, and social services agencies. Prior to public safety realignment, the CCP was formed to reduce recidivism to state prisons and then maintained as an advisory body.

Sonoma County designed its risk assessment system with the objective of helping judges make more consistent and transparent pretrial release or detention decisions. A pretrial risk assessment tool---the Sonoma County Pretrial Risk Assessment Tool (SPRAT)—was designed to predict the likelihood that individuals will commit pretrial misconduct. Then a policy framework was developed to facilitate interpretation of those risk predictions.

To create the SPRAT, researchers defined pretrial misconduct as a compound outcome of either arrest for a new crime or failing to appear in court and used existing criminal justice data to determine which factors predicted pretrial misconduct. The most predictive factors were criminal history, gang affiliation, homelessness, employment, and potentially violent mental health disorders.

To interpret the SPRAT risk predictions, CCP members collaborated with the courts to create a decision matrix that related risk level classifications and current offenses to pretrial release or detention decisions. The level of supervision increased with the SPRAT score and the severity of the offense. For instance, an individual who scored a 2 (of 4) on the SPRAT and who was booked for a petty theft could be released on own recognizance, while a person scoring a 3 who was arrested for domestic violence would be subject to stricter supervision.

Although Sonoma County has decided to transition from their SPRAT-based pretrial risk assessment system to one centered on the PSA, their experience provides valuable lessons for counties that may want to develop and evaluate their own pretrial risk assessment tools. In particular, the county evaluates the performance of their pretrial risk assessment system annually. The most recent report from 2018 examined overrides and "enhancements," which are conditions (e.g., threats to victims) that elevate risk classification levels above those predicted by the SPRAT. The analysis revealed that enhancements increased the number of people recommended for detention or enhanced supervision by 230 percent in 2018. Overrides by pretrial services officers also increased the number of people recommended for detention or enhanced supervision—but only by 13 percent—and mainly because the person was charged with a new crime. Judges also overrode SPRAT recommendations. Unlike pretrial services officers, they did so in both directions—some individuals who might have been detained were released and vice versa. Unfortunately, why judges departed from the SPRAT recommendations is unknown. Importantly, Sonoma County also examined racial inequity at six decision points in their pretrial risk assessment system, from whether an arrest resulted in a booking to whether a released defendant committed pretrial misconduct. Blacks were 5 times as likely as whites to be booked and 50 percent more likely to be recommended for detention or enhanced supervision before enhancements.

Takeaways

Convene a local stakeholder group.

Sonoma County repurposed an existing policymaking body to ensure that the relevant parties participated publicly in the development of its pretrial risk assessment system.

Be transparent.

The SPRAT was developed in a public forum, so the process used to develop the SPRAT was transparent. Likewise, the process through which individuals are classified is also transparent. How much each risk factor contributes to the overall risk score is explicitly stated. In addition, the decision matrix clearly illustrates how risk predictions are translated into pretrial release or detention decisions—and it is available online.

Avoid compound definitions of pretrial misconduct.

Creating a compound measure of pretrial misconduct reduced the transparency of the SPRAT. Compound outcomes are less transparent because it is unclear whether a person classified as high risk threatens public safety, is likely to miss a court date, or both. In addition, failure to appear and pretrial arrest are distinct outcomes with distinct predictors. Using the same variables to predict both outcomes simultaneously assumes that the predictors explain both outcomes similarly. Thus, the accuracy of the SPRAT may also have been negatively impacted.

Socioeconomic predictors may introduce inequity.

Of the SPRAT predictors, homelessness and mental health correlated most strongly with higher risk of pretrial misconduct. However, the Judicial Council has indicated that it may prohibit using these factors as "exclusions" because doing so can increase detention rates for people who are disadvantaged, rather than criminal. Before such factors are used in a risk prediction model, they can be tested to determine whether they propagate disadvantage.

Do not double-weight predictors.

Although the decision matrix transparently facilitates pretrial decisions, it double counts the same measure of criminal history by using it both to predict risk and as a component of the decision matrix. In addition, that weighting is often counteractive. For example, the SPRAT classifies individuals arrested for DUIs as very low risk of pretrial misconduct, but the decision matrix elevates an arrest for a DUI to a higher supervision status.

Revalidation is critical to assessing and addressing inequity in risk predictions.

Sonoma County's 2018 report highlights pretrial decision points where racial inequity can materialize. Their assessment indicated racial inequity at several of them. For the tool's performance, the most concerning are the inequities in pretrial risk predictions and pretrial release or detention recommendations. To address these inequities, the county can explore how alternative policies might exacerbate or ease them, as illustrated in the Center for Court Innovation case.

Regular evaluation is critical to understanding how systems perform over time.

Although pretrial release following a SPRAT assessment increased by 16 percent between 2016 and 2018, overrides and enhancements generally led to more restrictive pretrial release conditions. Enhancements are policies external to the pretrial risk assessment tool that affect how the system performs. If the county wants to release more people under less restrictive conditions, enhancement modifications may be required. Judicial downgrades present an opportunity to examine whether enhancements can be modified to allow release under certain circumstances.

Dressel and Farid Case Study

Transitioning to a pretrial risk assessment tool can create unique difficulties for counties that do not currently operate robust data collection systems. For example, using a pretrial risk assessment tool such as COMPAS, which uses 8 predictors that may be sourced from a core questionnaire that includes 137 items, may not initially be feasible for counties that currently collect only basic criminal justice and demographic information. Identifying additional predictors, hiring and training staff to collect them for each assessed person, and standardizing their use may be too steep a curve to overcome initially.

Dressel and Farid (2018) showed that more parsimonious and less resource intensive risk assessment tools can be developed. For counties with limited data resources seeking to transition to a risk based method of making pretrial release or detention decisions, the methods and models Dressel and Farid (2018) described may offer a more viable starting point for the local development a pretrial risk assessment tool. Using standard logistic regression methods for a sample of about 7,000 people, they created a risk prediction model using two predictors: age and total number of prior convictions.

Takeaways

Simpler risk prediction models can rival the accuracy of more complex models.

When Dressel and Farid (2018) compared their model to the COMPAS, they found that their tool correctly predicted outcomes 66.8 percent of the time, whereas the COMPAS correctly predicted outcomes 65.4 percent of the time. Although the overall accuracies of the two tools were similar, the types of errors they made were slightly different. The Dressel and Farid (2018) model incorrectly detained people at slightly higher rates than the COMPAS but also incorrectly released slightly fewer people.

Simpler risk prediction models can be similarly equitable across racial groups.

Dressel and Farid's (2018) two-predictor model was also similarly accurate for black and white individuals. Their model correctly predicted outcomes for whites 66.4 percent of the time compared to 67.0 percent for the COMPAS, and correctly predicted outcomes for 66.7 percent of blacks compared to 63.8 percent for the COMPAS.

More complicated pretrial risk assessment tools maintain certain advantages.

Pretrial risk prediction tools that use more information to predict risk tend to more accurately classify the most and least risky individuals because very high and very low risk classifications are made based on more robust information. Similarly, more complicated tools are able to make more accurate predictions when faced with individuals charged with less prevalent forms of criminal behavior, such as those charged with violent offenses.

ProPublica-COMPAS Case Study

In 2016 ProPublica published an article questioning the equity of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) pretrial risk assessment tool. According to ProPublica the COMPAS classified blacks as higher risk than whites even when they had similar criminal histories. Northpointe, the proprietor of COMPAS, argued that their tool was not inequitable or biased because the higher predicted risk for blacks accurately reflected the reality that blacks were more likely than whites to be arrested. Both parties were correct because each applied a different standard of equity (Mayson 2018).

Northpointe emphasized *predictive parity*, meaning a pretrial risk assessment tool should predict misconduct outcomes equally well for all individuals classified at a given risk level. For example, COMPAS expects about 60 percent of men of both races who are classified as high risk to be rearrested. ProPublica found that both black and white males classified by COMPAS as high risk were rearrested at about that rate. By this standard, the COMPAS pretrial risk assessment tool is not racially biased—the likelihood of correctly predicting rearrest is the same for both black and white men.

However, ProPublica applied a different standard of equity. *Statistical parity* expects individuals who experience particular pretrial misconduct outcomes to have been classified similarly. The COMPAS did not meet this standard. Among individuals who were not rearrested, 45 percent of blacks were classified as high risk, whereas only 23 percent of whites were. Similarly, among individuals who were rearrested, 48 percent of whites were classified as low risk, whereas only 28 percent of blacks were. By this standard, COMPAS is racially biased—more black men who are not rearrested are classified as high risk and fewer black men who are not rearrested are classified as low risk.

Takeaways

Policymakers need to consider the implications of failing to meet each standard of equity.

Failing to satisfy either standard of equity can have serious consequences for assessed individuals. Failing to achieve predictive parity means that risk classifications will be more accurate for one group than for the other—the predictions for whites are more likely to be correct than the predictions for blacks—which can lead to inappropriate pretrial detention or release for one group of people relative to the other. Failing to meet statistical parity can result in inequitable classification rates between groups—blacks are more likely than whites to be classified as high risk—which can lead to more pretrial detention in one group relative to the other.

County pretrial workgroups need to determine which standard of equity best promotes local policy objectives.

Simultaneously maximizing predictive parity and statistical parity is impossible because, as Northpointe noted, arrest rates vary for different groups of people. Although some balance between standards of equity can be achieved, policymakers will ultimately need to choose which standard to prioritize (Berk et al. 2018; Kleinberg et al. 2016; Mayson 2018). Which standard is prioritized should be decided publicly, so that the public understands the implications and tradeoffs of that decision.

Promoting either standard requires tradeoffs—specifically accuracy tradeoffs.

Increasing the equity—by either standard—of a pretrial risk assessment tool generally comes at the expense of reduced accuracy. For example, to increase the statistical parity of the COMPAS, whites could be classified as if

they were black, but doing so would mean detaining some whites who otherwise would be released—and thereby compromising their right to liberty. Conversely, blacks could be classified as if they were white, which would mean releasing some blacks who otherwise would be detained—and potentially threatening public safety. How to weigh these tradeoffs, again, should be considered in a public forum.

Criminal justice data reflect historical bias in the criminal justice system.

Arrest rates may differ for different groups of people because criminal justice data reflect historical bias in the criminal justice system. Historically blacks have been policed more heavily than whites, so it is unclear whether they are actually more likely to commit crime or just more likely to be arrested because they are monitored more closely. Yet pretrial risk assessments use these data to predict risk of pretrial misconduct as if there were not uncertainty in how they were created. Although there are limitations to how well such biases can be addressed, validation can help policymakers understand how accurate and equitable their pretrial risk assessments will be for different groups of people. From that baseline understanding, decisions can be made about how to interpret those risk predictions for all people and for protected classes of people, such as racial minorities. The Center for Court Innovation Case Study illustrates how racial bias can be mitigated—and at what cost.

Center for Court Innovation Case Study

Partly in response to the ProPublica-COMPAS debate, the Center for Court Innovation (CCI) determined whether their independently-developed pretrial risk assessment tool exhibits racial bias and whether it could be mitigated by policies that describe how to interpret and act on risk level classifications (Picard et al. 2019). CCI's recently released report illustrates how California's counties can validate their chosen pretrial risk assessment tools and evaluate their pretrial risk assessment systems to assess and mitigate racial inequity—and other potential inequities—that may emerge as pretrial risk assessment systems mature.

CCI tested their 9-item tool, which does not include race, but does include other demographic, criminal history, and current case information, using data from New York City. The tool classified individuals into five risk categories according to their predicted probability of being rearrested within two years: minimal, low, moderate, moderate-high, and high risk. When they initially tested their tool, CCI found that it classified individuals of all races—blacks, Latinos, and whites—with similar accuracy (AUC>=0.72).

However, CCI found evidence of racial inequity in risk level classifications, which can lead to racial inequity in pretrial detention rates. Blacks and Latinos were more likely than whites to be classified as moderate-high or high risk: 37 percent of blacks and 29 percent of Latinos were classified as moderate-high or high risk, but only 18 percent of whites were. If pretrial release or detention decisions were made based solely on these risk level classifications, fewer than 1 in 5 whites, but more than 1 in 3 blacks would be detained. Moreover, when they examined rearrest rates, CCI also found racial inequity in false positive rates. Blacks and Latinos classified as moderate high and high risk were more likely to *not* be rearrested than similarly classified whites: 23 percent of blacks and 17 percent of Latinos, but only 10 percent of whites were incorrectly classified.

CCI then tried to develop policies to mitigate these inequities by assessing how different alternatives would impact racial inequity in detention and false positive rates. First, CCI examined a policy of detaining only people classified as high risk. Under this alternative, both detention rates and false positive rates declined, but racial inequity remained. Detention rates were 22 percent for blacks, 10 percent for Latinos, and 16 percent for whites. False positive rates were 10 percent for blacks, 7 percent for Latinos, and 3 percent for whites. CCI then examined what would happen under a policy that limited detention to people classified as moderate-high and high risk who also were charged with a violent felony or domestic violence—by interpreting risk level classifications in concert with additional criminal history information. Relative to the first scenario, racial inequity was mitigated and detention rates declined—but false positive rates increased. Detention rates were 13 percent for blacks and whites and 14 percent for Latinos. False positives rates were 16 percent for blacks and Latinos and 14 percent for whites.

Takeaways

Pretrial risk assessment tools are likely to exhibit inequity—especially racial inequity.

People who have more prior arrests but are not more likely to commit crimes are more likely to be misclassified as high risk and are more likely to be needlessly detained as a result. Some demographic groups, especially racial minorities, are arrested at higher rates—even though they may not be more likely to commit crimes.

Validation helps counties determine the degree of racial inequity in risk prediction.

The CCI case illustrates how counties can determine the following across racial groups: (1) how a pretrial risk assessment tool will classify individuals; (2) to what degree those classifications are likely to be accurate; (3) and the consequences those classifications can have for pretrial release or detention decisions.

Pretrial risk assessment systems can be designed to promote equity.

After testing the performance of the risk assessment tool and finding racial inequity in potential detention rates and false positive rates, CCI created and tested policy alternatives to see whether they could reduce those inequities. Ultimately, they found a policy with the potential to promote racial equity by combining information from the pretrial risk assessment tool with an additional condition that recommends detention only when individuals have violent charges in their criminal histories.

Increased equity will generally come at the expense of reduced accuracy.

Risk assessment combined with detaining only potentially violent criminals increased equity in this case. But relative to a policy of only detaining the highest risk people, it comes at the cost of reduced accuracy. Although fewer people of all races are detained under the policy that creates more racial equity, false positive rates are higher for people of all races. Local pretrial policy objectives will determine whether this is an appropriate tradeoff, which is why those objectives need to be determined prior to validation.

Without robust pretrial risk assessment systems, pretrial decisions are likely to be more inequitable and inaccurate.

When CCI assumed that risk predictions would be translated directly into pretrial release or detention decisions, nearly 1 in 5 whites and more than 1 in 3 blacks would have been detained—and 1 in 4 blacks and 1 in 10 whites would have been incorrectly detained. CCI showed that detention rates could be reduced to less than 1 in 15 for all racial groups and that fewer than 1 in 15 people of all races would be incorrectly detained.

Pretrial risk assessment tools can be part of pretrial justice systems that lead to more transparent, consistent, accurate, and equitable pretrial release or detention decisions.

Non-proprietary pretrial risk assessment tools ensure that all people are evaluated in the same way, using the same criteria. How risk predictions are made is therefore transparent and consistent. The policies that govern how to interpret and act on those risk predictions should be similarly unambiguous and systematically applied. Under those conditions, pretrial release or detention decisions will be similarly transparent and consistent. As the CCI case illustrates, those policies can also be designed to ensure as much equity and accuracy as possible in pretrial release or detention decisions.

Appendix B. Predictors in Risk Assessment Tools

TABLE B1

Select characteristics of commercially available pretrial risk assessment tools currently used in California

	COMPAS PRRS-II	ORAS-PAT	VPRAI-R	PSA FTA	PSA NCA	PSA NVCA
Current Offense	Category representing most serious current charge		Current charge is felony drug, or theft			Current violent charge Current violent charge and age 20 or younger
Pending Charges	Number of pending charges or holds		Has pending charges	Has pending charges	Has pending charges	Has pending charges
Prior Pretrial Misconduct	Number of FTAs Number of times arrested or charged for new crimes during pretrial release	FTA warrants in the past 24 months: 0, 1, or more than 2	Has of two or more FTAs as an adult	Has FTA in the past two years Has FTA more than two years old	Has FTA in the past two years	
Prior Convictions			Has one or more past felony or misdemeanor convictions as an adult Has two or more violent convictions as an adult	Has prior felony or misdemeanor conviction	Has prior felony conviction Has prior misdemeanor conviction Has prior violent conviction	Has prior felony conviction Has prior misdemeanor conviction Has prior violent conviction
Prior Incarceration	Number of incarcerations that exceed 30 days	Has three or more prior incarcerations			Has prior sentence to incarceration	
Supervision Status			On community criminal justice supervision			
Age		Over or under age 33 at first arrest			Age at current arrest	
Employment	Employment status: full time, part time, unemployed, not in labor force	Employment status: full time, part time, unemployed	Employment status: employed, unemployed, student, caregiver, retiree, none			
Living Situation	Time in current neighborhood	Same residence for last six months				
Substance Use	Has history of drug use	Used illegal drugs in the last six months Drug use caused life problems in last six months	Has history of any drug use			

SOURCES: COMPAS Scale Documentation, Creation and Validation of the Ohio Risk Assessment Final Report, Virginia Pretrial Risk Assessment Tool - (VPRAI) Instruction Manual – Version 4.3, Public Safety Assessment Website

NOTES: FTA = FTA is failure to appear, NCA is new criminal act, and NVCA is new violent criminal act. The points assigned for NCA and FTA in the PSA risk assessment tool are totaled in two separate scales, whereas the total points for NVCA are converted to a binary "yes" or "no" outcome. The COMPAS pretrial release risk scale can be paired with the Violence Risk Scale to determine an individual's risk to the community.

Appendix C. Developing "State of the Art" Pretrial Risk Assessment Tools using Machine Learning: A Brief Introduction

Instead of validating existing pretrial risk assessment tools counties can develop and test their own. "State-of-the-art" risk assessment tools rely on algorithms (Berk 2019: 6). Algorithms are systematically applied decision rules. Algorithms can be very basic, generating a risk prediction using one or two pieces of information (e.g., Dressel and Farid 2018). Algorithms can also be very complex. For example, "decision trees," are processes that sequentially consider dozens or hundreds of variables to predict the likelihood of an outcome (Berk 2012, 2019; Kleinberg et al. 2017). Complex algorithms, including the decision trees that have been used to predict pretrial outcomes, are identified using "machine learning" techniques, meaning a computer is supplied with data and directed to predict an outcome using a specified methodology. The computer adaptively creates and revises the algorithm as it incorporates more data. Like the comparison between clinical and actuarial assessments, pretrial risk assessment tools based on machine learning algorithms have been shown to be more accurate than those based on statistical models, such as logistic regression (Berk et al. 2014; Kleinberg et al 2017).[2]

Using Machine Learning to Develop Pretrial Risk Assessment Tools

Developing a pretrial risk assessment tool begins with at least three decisions. First, local policy objectives must be established. Second, the outcome to be predicted must be defined. Third, the data used to predict the outcome must be selected to reflect the local population and policy environment. Each of these steps is described in the main report, so we only briefly summarize them here. After those decisions are made, the processes of developing and validating the risk prediction model that will undergird the pretrial risk assessment tool can begin.

Define Pretrial Policy Objectives

Pretrial policy objectives operationalize the goals of pretrial justice, which include maximizing individual liberty, public safety, court appearances, and equity. How counties operationalize these goals will influence the design of the pretrial risk assessment system, from defining pretrial misconduct to making pretrial release or detention decisions. When county pretrial workgroups convene, they should begin by defining these objectives.

Precisely Define the Pretrial Misconduct Outcome to Predict

Risk prediction models can only predict well-defined outcomes and they are "*exceedingly sensitive"* to the choice of outcome (Kleinberg et al. 2019: 5, emphasis in original). As described in the main report, counties will need to precisely define the pretrial misconduct outcomes they want to predict. Current legal scholarship indicates that individuals should be detained only to prevent serious violent crimes during the pretrial period (Mayson 2019; PDRW 2017).

[2] The distinction between machine learning algorithms and statistical models is not always clear. A useful distinction may be that machine learning algorithms typically impose less structure on the data than statistical methods because they do not assume an underlying model, whereas statistical methods typically do (Berk 2019). However, there are statistical methods that also do not impose structure on the data.

Rare Outcomes

Accurately predicting rare outcomes is a fundamental challenge for all risk prediction models. The best available data indicates that violent crimes are committed during the pretrial period very rarely. For example, between 1990 and 2009 only 1.4 percent of felony defendants in California were arrested for a violent felony during the pretrial period (Tafoya 2015). As a result, those forms of pretrial misconduct that pose the greatest threat to public safety are also the most difficult to predict accurately.

Rare outcomes pose two key problems for algorithmic risk assessment tools: they limit the number of similar cases that can be used to train the model and they make calibrating the tool to appropriately reflect the rarity of the outcome challenging. We discussed the first problem at length in the main report: for a pretrial risk assessment tool to make accurate risk predictions for future arrested individuals, it must have a robust sample of past similar arrestees. Rare outcomes like violent felonies make for a small sample—particularly in less populated counties—and may not provide enough observations to create unique training and testing datasets.

The second calibration problem is subtler. Because violent felonies occur so rarely, it is difficult for a tool to both assign a probability that reflects their rarity and be appropriately sensitive to their occurrence. For example, if a county has a violent felony arrest rate of 300 per 100,000 residents, then the risk prediction model should (at most) predict that 3 percent of individuals will commit a violent felony while on pretrial release. Therefore, the predicted probability of pretrial violence should be near zero for most assessed individuals. And those with non-zero predicted probabilities of pretrial violence should still have low probabilities overall. Setting a threshold to separate low from very low probabilities of pretrial violence will tend to lead to either too many (i.e., over-sensitivity) or not enough (i.e., under-sensitivity) people predicted to commit a violent felony.

In addition, traditional performance metrics like accuracy, which are intended to reflect how well tools predict risk, can provide deceptive information (Hester 2019). Referring to our previous example, if the risk assessment tool never predicts anyone will commit a violent felony, it will still be accurate 97 percent of the time because it will make incorrect predictions only for the 3 percent of individuals who do commit violent felonies. Yet the tool will fail to predict violence for 100 percent of the instances in which it occurs. Similarly, the tool could grossly over predict the number of people likely to commit a violent felony, and still result in a very high accuracy. As a result, researchers and practitioners responsible for validating the performance of risk assessment tools should carefully examine the different types of errors the tool makes in predicting rare outcomes, rather than relying on more general diagnostics that reflect the overall performance of the tool. We describe how to do this in Technical Appendix D.

Collect Representative Data to Develop and Test a Risk Prediction Model

Pretrial risk assessment tools unavoidably use information from the past to predict the future. When gathering data to develop and test tools, counties should try to gather past data that best represents the current local policy landscape and the current local pretrial population. As described in the main report, important considerations include whether there have been substantial demographic shifts or shifts in the policy environment that may affect pretrial misconduct outcomes (e.g., Bird et al. 2016).

To build a representative dataset, information on all pretrial release or detention decisions and pretrial misconduct outcomes in a county over a relevant time period should be gathered, as should additional systematically collected data that can be used to make pretrial risk assessments (e.g., demographic, criminal history, and socioeconomic information). Counties should gather as much information as possible and include it in the dataset—no predictors should be excluded a priori (e.g., due to equity concerns). Machine learning models perform better when more data is available to them. Whether including particular predictors compromises equity can be evaluated later.

Developing and Testing a Risk Prediction Model

In a machine learning framework, developing a risk assessment tool essentially amounts to developing (and choosing) the best performing risk prediction model and then testing it. Both developing (i.e., training) a risk prediction model and validating (i.e., training) it require unique samples or subsets of the representative dataset.[3] How much data—meaning how many observations—the dataset contains therefore determines whether and which machine learning techniques can be used to develop the tool. In jurisdictions with larger volumes of pretrial release or detention decisions and misconduct outcomes (e.g., 10,000 or more), machine learning techniques that rely on "big data" are feasible (e.g., Kleinberg et al. 2017). In jurisdictions with fewer pretrial release or detention decisions and pretrial misconduct outcomes (e.g., 1,000), the methods differ, but may still exploit recently developed machine learning techniques (e.g., Berk et al. 2014).

Large Sample Machine Learning Techniques

To apply machine learning techniques in large samples, the representative dataset will be divided into a minimum of two subsets. The first subset is used for "training." In the training stage, the data are used to incrementally improve upon a risk prediction model until a version of the model that best predicts the desired pretrial misconduct outcome while also satisfying the local policy objectives is identified. The second subset is used to test the chosen risk prediction model, meaning to reassess its performance using data it has not yet seen. This process is akin to the validation process that we described in the main report. A third "verification" subset is often desirable (but not strictly required) because it enables a second independent test of the chosen risk prediction model (e.g., Kleinberg et al. 2017; Berk 2012, 2019).[4]

Small Sample Machine Learning Techniques

Berk et al. (2014) developed the only machine learning process for small samples (n~1500) of which we are aware. Their process, which is available as an R package, relies on kernel methods and requires three unique subsets of the data. Training data are used to identify several promising risk prediction models. A second "specification" dataset is used to identify the best performing risk prediction model from among the promising models. Finally, testing data is used to assess the performance of the chosen risk prediction model on new data.

Limitations of Machine Learning: Transparency and Complexity

Although machine learning algorithms often outperform simpler statistical models, they are less transparent in how they reach their predictions. For statistical models, analysts can directly examine the predictors and the risk prediction model to understand how a risk prediction will be reached. That is not possible with machine learning algorithms. Machine learning algorithms can be adjusted. But to understand what happens when a machine learning risk prediction model runs, it must be run (Kleinberg et al. 2019).

Finally, county agencies that are already overwhelmed with administrative tasks, policy development, and policy evaluation may find it difficult to allocate the time and resources necessary to learn and applying machine learning techniques to the development of pretrial risk assessment tools. Counties may therefore find it fruitful to collaborate with academic institutions or research consulting firms to develop such tools.

[3] Ideally, each subset will include unique observations: the available data will be divided so that each observation appears in only one of the subsets. Alternatively, random samples can be taken from the available data. In the later scenario, each subset will be unique, but some observations will be repeated across the subsets.
[4] Berk (2012) steps through these processes, provides some examples of machine learning code, and provides additional references for deeper learning.

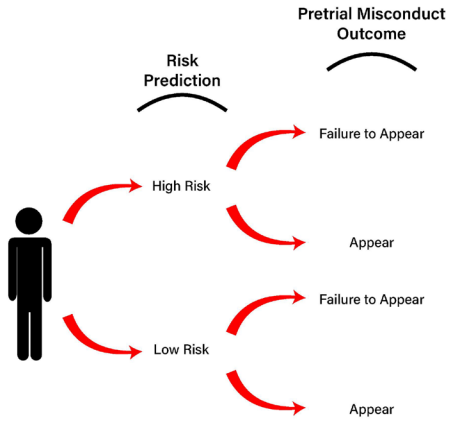
Appendix D. Performance of Risk Assessment Tools

Risk assessment tools are validated by considering different aspects of their performance, meaning how well the predictions made by the risk assessment tool conform to future behavior on the part of the assessed individuals. In the sections that follow, we present the most common performance measures and illustrate how they are derived. We then explain the intuition behind each measure.

Relationships between Risk Predictions and Pretrial Misconduct Outcomes

When there are two risk predictions, high risk or low risk, and two pretrial misconduct outcomes, failure to appear (FTA) or appear, the paths from risk prediction to pretrial misconduct outcomes can be depicted as in Figure D1. Figure D1 can then be translated into what is called a confusion table, as depicted in Table D1. A confusion table relates the risk predictions made by risk assessment tools to the behavior observed after the prediction was made.

FIGURE D1.
Relating risk predictions to pretrial misconduct outcomes



SOURCE: Author illustration

The four cells at the center of the confusion table reflect the four potential relationships between pretrial risk predictions and pretrial misconduct outcomes shown in Figure D1: two ways of making correct risk predictions and two ways of making incorrect risk predictions. In this framework, "true" means correct, "false" means incorrect, "positive" indicates the predicted behavior (in our example, failure to appear), and "negative" indicates the opposite of the predicted behavior (in our example, appear).

TABLE D1

Confusion table that represents the relationship between predicted risk and actual behavior on pretrial release

		Actual Pretrial Misconduct Outcome		Equity (Outcome Oriented)
Pretrial Risk Prediction		**Fail to Appear**	**Appear**	Statistical Parity
	High Risk	True Positive TP	False Positive FP	False Positive Rate FP/(FP+TN)
	Low Risk	False Negative FN	True Negative TN	False Negative Rate FN/(FN+TP)
Equity (Prediction Oriented)	Predictive Parity	False Discovery Rate FP/(FP+TP)	False Omission Rate FN/(FN+TN)	
Accuracy	Accuracy	(TP+TN)/(TP+FP+FN+TN)		
	Calibration	Percent Appear	Percent Low Risk	

SOURCE: Author illustration

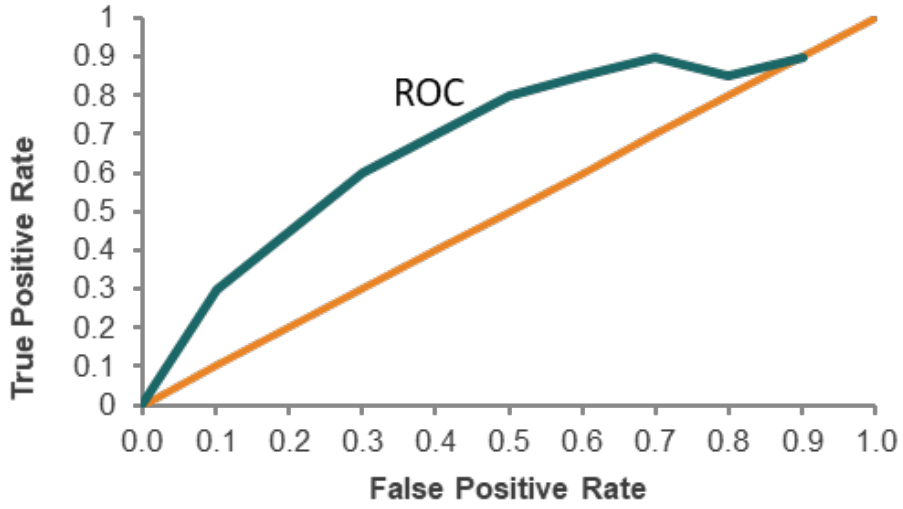
In the sections that follow, we first review the most common overall performance metric, the area under the curve. We then review more fine-grained indicators of performance. We introduce two different perspectives on the performance of risk assessment tools and show why they matter for the measurement of performance. The same tool can be said to perform well or poorly, depending on the perspective adopted. Finally, we highlight the consequences for accuracy and equity of measuring performance from each perspective.

Overall Performance: Area under the Curve

The most common performance measure is called the *area under the curve* (AUC). The "curve" is the *receiver operating characteristic* (ROC) curve, which is a plot of the true positive rate (on the y-axis) as a function of the false positive rate (on the x-axis), as depicted by the blue line in Figure D2. The ROC curve visually represents the tradeoff between assigning a high risk classification to individuals likely to commit pretrial misconduct (i.e., making a correct prediction) and assigning a high risk classification those who are unlikely to commit pretrial misconduct (i.e., making an incorrect prediction). The AUC measures the distance between the ROC and an idealized relationship between the correct and incorrect predictions, which is represented by the orange line in Figure D2. This line represents a 1:1 ratio between correct and incorrect predictions.

Intuitively, a risk prediction model performs better by making more correct than incorrect predictions: the ratio between correct and incorrect predictions is greater than 1:1. When that occurs, the ROC line will lie above the idealized line, as shown in Figure D2. The greater the distance between the ROC line and the idealized line, the more true positives the risk prediction model assigns relative to false positives. Taking the integral of the ROC relative to the idealized line produces the AUC.

FIGURE D2
Hypothetical area under the curve plot



SOURCE: Author illustration

AUCs can range from 0.0 to 1.0, with 1.0 indicating perfectly accurate prediction and 0.0 indicating perfectly inaccurate prediction. An AUC of 0.5 means that the tool has a 50 percent chance of distinguishing a high risk person from a low risk person—no better than flipping a coin. An AUC of 1.0 means that the tool has a 100 percent chance of distinguishing a high risk person from a low risk person. Generally, an AUC value greater than 0.7 signals that the risk prediction model makes adequately accurate predictions, whereas values below 0.6 suggest that it does not.

Perspectives on Performance

Using the four basic relationships at the heart of the confusion table, we can examine the performance of a risk assessment tool from two perspectives. Within each perspective both correct and incorrect predictions are possible. However, analysts typically evaluate the performance of risk assessment tools in terms of the false or incorrect predictions, rather than the true or correct predictions. In other words, they want to understand prediction errors so that they can be corrected. The key fact to recognize is that false positives and false negatives can be measured in two ways, from two perspectives.

Prediction-Oriented Perspective

A *prediction-oriented perspective* looks forward from predictions to outcomes and asks: at what rate did the risk predictions fail to materialize? At what rate did high risk people appear; and at what rate did low risk people fail to appear? The former is called the *false discovery rate* (mathematically: FP/FP+TP). The latter is called the *false omission rate* (mathematically: FN/FN+TN).[5]

[5] These false rates have corresponding true rates: at what rate did the risk predictions materialize as actual outcomes? For two-by-two confusion tables, the true rates oppose the false rates.

Outcome-Oriented Perspective

Alternatively, an *outcome-oriented perspective* looks backward from outcomes to predictions and asks: at what rate were the outcomes predicted incorrectly? At what rate were the people who failed to appear predicted to appear; and at what rate were the people who appeared predicted to fail to appear? The former is called the *false negative rate* (mathematically: FN/FN+TP). The latter is called the *false positive rate* (mathematically: FP/FP+TN).[6]

Defining Accuracy and Calibration

Accuracy in risk assessment is most often defined as the proportion of correct predictions: the number of true positives plus the number of true negatives, divided by the total number of predictions (mathematically: (TP+TN)/(TP+TN+FP+FN)).[7] Defined in this way, accuracy depends on each of the four core relationships between risk predictions and actual behavior. This is also a very intuitive definition of accuracy.

However, accuracy can be defined in more than one way. Another definition of accuracy has been called "calibration" (Kleinberg et al. 2016: 4). Calibration asks whether the proportion of people predicted to appear matches the proportion of people who actually appear regardless of whether those predictions are correct or incorrect (mathematically: (FP+TN)/(TP+TN+FP+FN)).

To see why calibration is an important alternative measure of accuracy, consider a population in which only 80 percent of people appear but the tool predicted that 50 percent are at high risk for failing to appear. The performance of the tool is immediately called into question because it predicts that far more people will fail to appear than actually do fail to appear. Thus, calibration is an important initial test of the performance of a risk assessment tool. It requires that risk scores "mean what they claim to mean" even if the predictions are sometimes incorrect (Kleinberg et al. 2016: 4).

Predictive Parity and Statistical Parity

In Technical Appendix E, we discuss seven standards of equity. Here, we discuss in more detail the two we highlighted in the main report. *Statistical parity* adopts an outcome-oriented perspective by looking backward from an outcome to ask how many people in each group were predicted to experience it. *Predictive parity* adopts a prediction-oriented perspective by looking forward from a prediction to ask how many people in each group experienced the outcome.

A tool achieves statistical parity when the false positive rate and the false negative rate are the same for both groups of people.[8] More intuitively, the percentage of people who appeared and who were initially classified high risk should be the same in both groups. Likewise, the percentage of people who failed to appear and who were initially classified as low risk must be the same in both groups.

Predictive parity requires the false discovery rate and the false omission rate to be the same for both groups of people. More intuitively, the percentage of people classified as high risk and who go on to appear must be the same in both groups. Likewise, the percentage of people classified low risk and who go on to fail to appear must be the same in both groups.[9]

---

[6] Similarly these false rates have corresponding true rates: at what rate were the actual outcomes predicted?

[7] The complementary measure to accuracy is the misclassification rate, defined as proportion incorrect predictions (mathematically: FP+FN/TP+TN+FP+FN).

[8] Berk et al. (2018) refer to this as "conditional procedure accuracy equality." We chose a term that references more commonly used terms in the broader risk assessment literature.

[9] Berk et al. (2018) refer to this as "conditional use accuracy equality." We chose to follow Chouldechova's (2017) lead because her terminology references the common definitions of the true composite terms: positive predictive value (TP/(TP+FP)) and negative predictive value (TN/(TN+FN)).

Like accuracy, statistical parity and predictive parity rely on each of the four relationships between risk predictions and actual behavior. Intuitively, these relationships suggest that there will be tradeoffs between accuracy, statistical parity, and predictive parity. To demonstrate why those tradeoffs are inevitable in real-world situations, we introduce two more concepts: base rates and error weights.

Accuracy, Equity, and Base Rates of Pretrial Misconduct

Base rates refer to the underlying probability that an outcome will occur in a population or in subsets of that population. Different subsets of a population (e.g., groups delineated by race, age, or socioeconomic status) do not necessarily have equal probability of experiencing pretrial misconduct outcomes. Their base rates of failing to appear or committing a crime during pretrial release differ.

The potential for underlying variation in base rates of experiencing pretrial misconduct outcomes complicates the notions of equity and accuracy that we have been discussing. To understand why consider the tables in Panels A, B, and C of Figure D3. The tables in each panel are laid out as in Table D1, but with additional cells that indicate the total number of assessed individuals, the number of individuals who were predicted high and low risk, and the number of individuals who failed to appear and appeared.

In Panel A, we present an idealized hypothetical situation that allows us to discuss some features of risk assessment tool performance that can help counties compare how risk assessment tools perform for different population subgroups. First, notice that the tool that produced these results is *calibrated*: the base rate of appearing in Group 1 is 50 percent and half of the people in Group 1 are classified as low risk. Second, notice that the false positive and false negative rates are the same. Moreover the number of false positive and false negatives is the same, suggesting that policymakers value both false positives and false negatives similarly. This is rarely the case in real-world applications. Finally, note that the false discovery and false omission rates are also the same. Again, this rarely occurs in real-world situations.

In Panel B, we present the performance of the same risk assessment tool for Group 2, another hypothetical situation intended to illustrate how base rates can impact notions of equity between groups of people. Base rates of failing to appear in Group 2 (67 percent) are higher than in Group 1 (50 percent). Mathematically, this is achieved simply by multiplying the rightmost column by 2, which means there are 3000 people in Group 2, whereas there were 2000 people in Group 1. Note what happens to the performance measures. False positive and false negative rates remain equal and, in fact, are the same as for Group 1. Statistical parity is also achieved. However, predictive parity is compromised. More Group 1 members appear (50 percent versus 33 percent) and fewer fail to appear (50 percent versus 67 percent) than Group 2 members. Yet fewer Group 1 members are classified as low risk (40 percent versus 57 percent) and more are classified as high risk (40 percent versus 25 percent) than Group 2 members. Calibration is partly to blame: the tool predicts that 47 percent of Group 2 members will appear when in fact only 33 percent will. The calibration problem can be fixed. However, as Panel C illustrates, fixing the calibration problem increases the accuracy of the predictions for Group 2 but does not necessarily increase equity relative to Group 1.

In Panel C, some Group 2 members who eventually fail to appear are shifted from the low risk level classification to the high risk level classification. A shift like this seems appropriate and, intuitively might be accomplished by moving the rightmost line in Figure 2 in the main report to the left. As Panel C illustrates, this shift achieves calibration for Group 2. Thirty-three percent of Group 2 members appear and 33 percent of Group 2 members are classified as low risk. However, the predictive parity gains that accompany this shift come at the expense of statistical parity. False omission rates are the same for both groups and the false discovery rate for Group 2 is

closer to that of Group 1 than it had been. But the false negative rate is lower for Group 2 than it is for Group 1, even though false negatives and false positives are again valued equally in both groups.

FIGURE D3

How different base rates of failing to appear can impact policy decisions related to accuracy and equity

Panel A: Risk Assessment Performance for Group 1

		Actual Behavior			
		Fail to Appear	Appear	N	Statistical Parity
Predicted Risk	High Risk	600	400	1000	0.40
	Low Risk	400	600	1000	0.40
	N	1000	1000	2000	
	Predictive Parity	0.40	0.40		
	Accuracy	0.60			
	Calibration	0.50	0.50		

Panel B: Risk Assessment Performance for Group 2 (Only Base Rates Differ)

		Actual Behavior			
		Fail to Appear	Appear	N	Statistical Parity
Predicted Risk	High Risk	1200	400	1600	0.40
	Low Risk	800	600	1400	0.40
	N	2000	1000	3000	
	Predictive Parity	0.25	0.57		
	Accuracy	0.60			
	Calibration	0.33	0.47		

Panel C: Risk Assessment Performance Calibrated for Group 2

		Actual Behavior			
		Fail to Appear	Appear	N	Statistical Parity
Predicted Risk	High Risk	1600	400	2000	0.40
	Low Risk	400	600	1000	0.20
	N	2000	1000	3000	
	Predictive Parity	0.20	0.40		
	Accuracy	0.73			
	Calibration	0.33	0.33		

SOURCE: Adapted from Berk et al. (2018)

Finally, a different kind of equity also seems to be compromised: the overall accuracy of the tool improved from 60 percent to 73 percent for Group 2, which far exceeds the accuracy of the tool for Group 1 (60 percent). If accuracy is greater for Group 2 than for Group 1, it means that the members of Group 1 are more likely to be treated inequitably because the classifications applied to them are more likely to be incorrect.

Figure D3 illustrates a proven "impossibility theorem" (Berk et al. 2018: 17; Kleinberg et al. 2016). In the absence of perfect prediction, if base rates are unequal it is impossible to maximize *both* statistical parity, and predictive parity simultaneously. Although they can be better balanced as the shifts between panels illustrate, policymakers must choose which to sacrifice in service to the other.

The Cost of Making Mistakes: Accuracy, Equity, Liberty, and Safety

Variation in base rates is not the only factor policymakers need to consider as they decide how to predict risk and translate those risk predictions into pretrial release or detention decisions. Risk classification choices assign value to prediction errors—false negatives and false positives—which reflect choices about how individual liberty is valued in relation to public safety.

To begin to understand this, consider Panels A and C of Figure D3. In Panel C, the *cost ratio*, meaning the ratio of false negatives to false positives, is 1:1. The risk prediction model allows the same number of the different types of errors. In Panel B, however, the cost ratio is 2:1. The risk prediction model allows twice as many false negative as false positives. The implication in Panel A is that public safety and individuals' right to liberty are valued equally. In Panel C, the implication is that public safety is half as valuable as individuals' right to liberty, because false negatives are most likely to impact public safety, whereas false positives are most likely to impact individuals' right to liberty.

The notion of "valuing errors" might seem overly technical. But people intuitively understand and implicitly "value" false negatives and false positives. If a person classified as low risk is released and commits a new crime, the victim, the victim's family, and the local community primarily bear the consequences of the false negative—public safety is compromised. Likewise, if a person classified as high risk is detained, but would not have committed a new crime, he, his family, and his community primarily bear the consequences of the false positive—the individual right to liberty is compromised. Thus, the exercise of placing value on errors and estimating the consequences of that valuation can help policymakers better understand the tradeoffs inherent in predicting risk and making decisions based on those predictions that impact their constituents' liberty and safety.

Appendix E. Equity Standards in Pretrial Risk Assessment

Equity can be understood as a measure of whether a risk assessment tool treats different types of people equally. In the pretrial literature discussions of equity have largely centered on race but can also be extended other classes of people (e.g., gender, socioeconomic status, and health). Below we review seven standards of equity, which are primarily referred to as standards of "fairness" in the academic literature, to provide policymakers with a sense of the tradeoffs they may face when deciding which to promote.

Predictive Parity

Predictive parity requires the positive predictive value (precision) and the negative predictive value to be the same for both groups. Predictive parity also implies that the false discovery and false omission rates should be the same for both groups, which we distinguish with "1" and "2" subscripts in the following equations (Berk et al. 2018).

$$\frac{FP_1}{TP_1 + FP_1} = \frac{FP_2}{TP_2 + FP_2}$$

$$and$$

$$\frac{FN_1}{TN_1 + FN_1} = \frac{FN_2}{TN_2 + FN_2}$$

Statistical Parity

Statistical parity requires that the false positive and false negative rates be the same for the two groups. Statistical parity also implies sensitivity-specificity parity, meaning that the true positive rate (sensitivity) and the true negative rate (specificity) should be the same for both groups.

$$\frac{FP_1}{FP_1 + TN_1} = \frac{FP_2}{FP_2 + TN_2}$$

$$and$$

$$\frac{FN_1}{FN_1 + TP_1} = \frac{FN_2}{FN_2 + TP_2}$$

Accuracy Equality

Accuracy equality requires the proportion of correct predictions to be the same in each group. In other words, the accuracy of prediction should be the same for both groups.

$$\frac{TP_1 + TN_1}{TP_1 + TN_1 + FP_1 + FN_1} = \frac{TP_2 + TN_2}{TP_2 + TN_2 + FP_2 + FN_2}$$

Demographic Parity

Demographic parity requires the proportion of people predicted to be high risk to be the same for both groups.

$$\frac{TP_1 + FP_1}{TP_1 + TN_1 + FP_1 + FN_1} = \frac{TP_2 + FP_2}{TP_2 + TN_2 + FP_2 + FN_2}$$

$$and$$

$$\frac{TN_1 + FN_1}{TP_1 + TN_1 + FP_1 + FN_1} = \frac{TN_2 + FN_2}{TP_2 + TN_2 + FP_2 + FN_2}$$

Treatment Parity

Treatment parity requires the "cost ratio" of false negatives to false positives be the same for both groups.

$$\frac{FN_1}{FP_1} = \frac{FN_2}{FP_2}$$

Calibration Parity

Although they do not include it among their definitions of equity, Berk et al. (2018) adopt Kleinberg et al.'s (2016) definition of calibration as correctly predicting the probability of experiencing an outcome regardless of prediction errors. They argue that calibration parity is an important definition of equity, so we include it here.

$$\frac{FP_1 + TN_1}{TP_1 + TN_1 + FP_1 + FN_1} = \frac{FP_2 + TN_2}{TP_2 + TN_2 + FP_2 + FN_2}$$

Total Equity

Total equity occurs when all parity measures are achieved. This occurs only in the trivial and not realistic case in which different groups have identical base rates.

Appendix F. Example Decision Matrixes and Decision Tree

FIGURE F1

New Jersey's Pretrial Release Recommendation Decision Making Framework

Pretrial Release Recommendation Decision Making Framework (DMF)
[March 2018]

DMF MATRIX

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	Risk Level Green – Recommendation ROR	Risk Level Green – Recommendation ROR				
FTA 2	Risk Level Green – Recommendation ROR	Risk Level Green – Recommendation ROR	Risk Level Light Green – Recommendation PML 1	Risk Level Yellow – Recommendation PML 2	Risk Level Light Orange – Recommendation PML 3	
FTA 3		Risk Level Light Green – Recommendation PML 1	Risk Level Light Green – Recommendation PML 1	Risk Level Yellow – Recommendation PML 2	Risk Level Light Orange – Recommendation PML 3	Risk Level Red – No Release Recommended
FTA 4		Risk Level Light Green – Recommendation PML 1	Risk Level Light Green – Recommendation PML 1	Risk Level Yellow – Recommendation PML 2	Risk Level Light Orange – Recommendation PML 3	Risk Level Red – No Release Recommended
FTA 5		Risk Level Yellow– Recommendation PML 2	Risk Level Yellow – Recommendation PML 2	Risk Level Light Orange – Recommendation PML 3	Risk Level Dark Orange – Recommendation PML 3 + EM/HD	Risk Level Red – No Release Recommended
FTA 6				Risk Level Red – No Release Recommended	Risk Level Red – No Release Recommended	Risk Level Red – No Release Recommended

SOURCE: https://njcourts.gov/courts/assets/criminal/decmakframwork.pdf

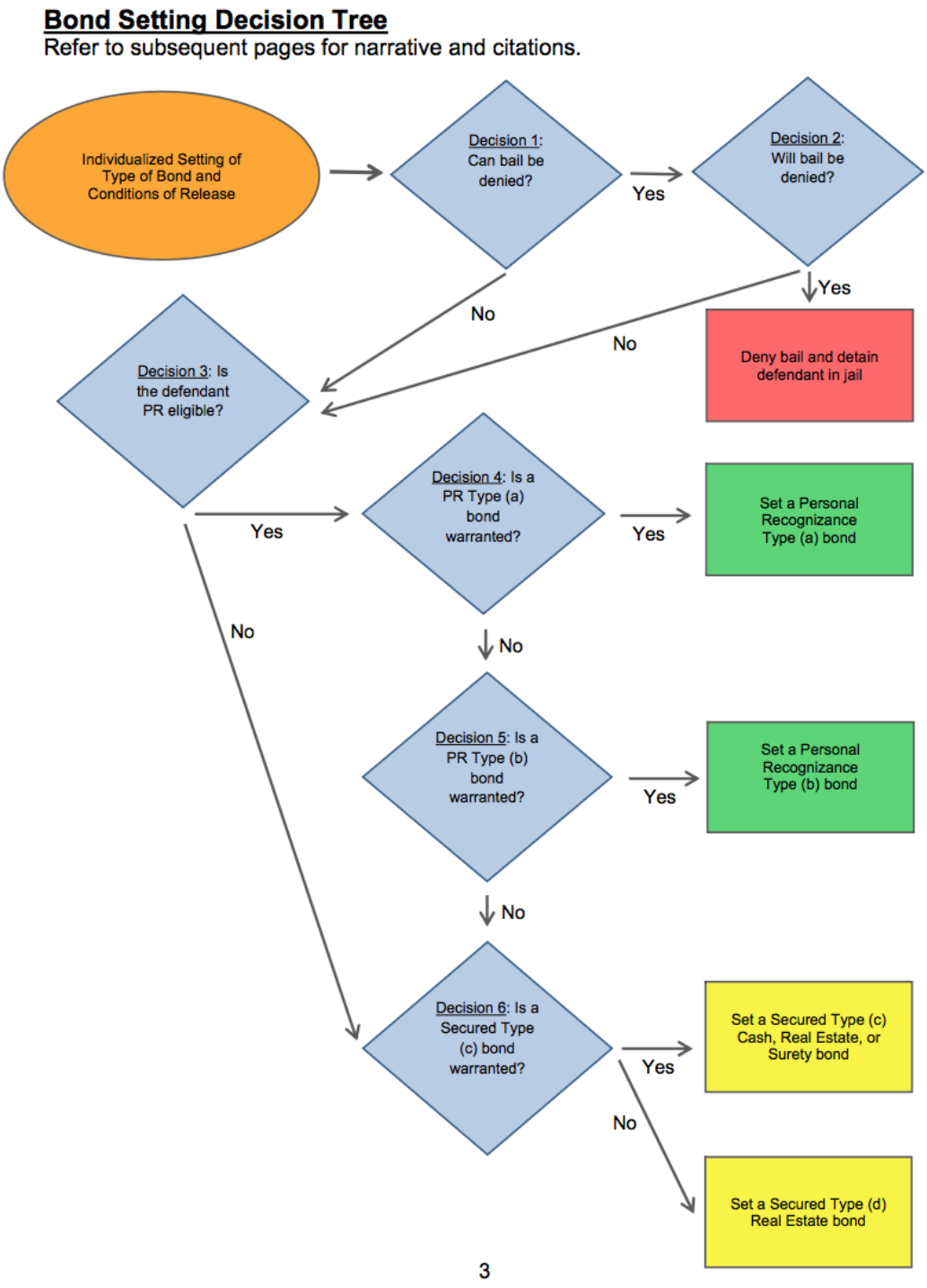
FIGURE F2

Example Decision Matrix from Chief Probation Officers of California and the Pretrial Justice Institute

	MOST SERIOUS CHARGE					
PRETRIAL RISK CATEGORY	LESS SERIOUS MISDEMEANOR	MORE SERIOUS MISDEMEANOR	LESS SERIOUS OR NON-VIOLENT FELONY	DRIVING UNDER THE INFLUENCE	DOMESTIC VIOLENCE	SERIOUS OR VIOLENT FELONY
LOWER	Recognizance Release with Court Reminder	Recognizance Release with Court Reminder	Recognizance Release with Court Reminder	Recognizance Release with Basic Supervision	Recognizance Release with Basic Supervision	Recognizance Release with Enhanced Supervision (if Released); or Detained
MEDIUM	Recognizance Release with Basic Supervision	Recognizance Release with Basic Supervision	Recognizance Release with Basic Supervision	Recognizance Release with Enhanced Supervision	Recognizance Release with Enhanced Supervision	Recognizance Release with Enhanced Supervision (if Released); or Detained
HIGHER	Recognizance Release with Basic Supervision	Recognizance Release with Enhanced Supervision	Recognizance Release with Enhanced Supervision	Recognizance Release with Enhanced Supervision (if Released); or Detained	Recognizance Release with Enhanced Supervision (if Released); or Detained	Recognizance Release with Enhanced Supervision (if Released); or Detained

SOURCE: CPOC (2019)

FIGURE F3
Colorado's Bond Setting Decision Tree



SOURCE: Jones and Schnake (2013)

PPIC

PUBLIC POLICY
INSTITUTE OF CALIFORNIA

25 YEARS

The Public Policy Institute of California is dedicated to informing and improving public policy in California through independent, objective, nonpartisan research.

Public Policy Institute of California
500 Washington Street, Suite 600
San Francisco, CA 94111
T: 415.291.4400
F: 415.291.4401
PPIC.ORG

PPIC Sacramento Center
Senator Office Building
1121 L Street, Suite 801
Sacramento, CA 95814
T: 916.440.1120
F: 916.440.1121