The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment

Sharad Goel
Stanford University

Ravi Shroff
New York University

Jennifer Skeem
University of California—Berkeley

Christopher Slobogin
Vanderbilt University

Jurisdictions across the country, including the federal government through its recently enacted First Step Act, have begun using statistical algorithms (also called "instruments") to help determine an arrestee's or an offender's risk of reoffending. Most instruments are relatively simple tools that assign the individual to a risk category representing the probability of recidivism if not detained. Some algorithms aim to provide information not only about risk assessment but also about risk management, or the type of intervention that might most effectively reduce risk.

These risk assessment instruments (RAIs) might be used at a number of points in the criminal process (Christin, Rosenblat & Boyd, 2015). They may be used at the front-end by judges to impose a sentence after conviction, at the back-end by parole boards to make decisions about prison release, or in between these two points by correctional authorities determining the optimal security and service arrangements for an offender. At the pretrial stage, RAIs might come into play at the time of the bail or pretrial detention determination by a judge, which usually takes place shortly after arrest. As a general matter, judges, parole boards and correctional officials have discretion as to how much weight to give the outputs of such instruments.

Prior to the advent of RAIs, legal decision makers called upon to evaluate an offender's risk usually relied on the opinions of mental health professionals, probation officer assessments, or their own seat-of-the pants analysis. This type of judgment is often called "clinical" prediction—to distinguish it from "actuarial", statistically based prediction—and it is still the basis of the post-conviction and pretrial decision-making process in many jurisdictions.

The increased use of RAIs in the criminal justice system has given rise to several criticisms. RAIs are said to be no more accurate than clinical assessments, racially biased, lacking in transparency and, because of their quantitative nature, dehumanizing. This chapter critically examines a number of these concerns. It also highlights how the law has, and should, respond to these issues.

PART I: ACCURACY

Risk assessment instruments are purpose-built to predict reoffending. The rationale for using RAIs to inform decision making in the criminal justice system is that RAIs can predict reoffending more consistently, transparently, and accurately than unaided human judgment—i.e., the intuitive opinion of a judge, probation officer, clinician, or other professional. Recently, however, this rationale has come under fire. Despite more than a half-century of research indicating that decisions are more accurate when professional judgment is structured or replaced by algorithms, authors of a recent study published in Science Advances claim they found that a widely used algorithm "is no more accurate or fair than predictions made by people with little or no criminal justice expertise" (Dressel & Farid, 2018). In this section, we resolve this apparent contradiction while outlining the current state of science on the relative accuracy of RAIs in assessing justice-involved people's risk of reoffending.

Algorithms typically outperform unguided human predictions

Algorithms typically outperform human judgment in predicting many outcomes, including recidivism. In a classic book that shaped the nascent risk assessment field, psychologist Paul Meehl (1954) distinguished two methods of predicting human behavior: information could be combined in a professional's head using personal judgment (the clinical method) or combined using "empirically established relations between data and the condition or event of interest" (the actuarial method; Dawes, Faust & Meehl, 1989, p. 1668). Today, meta-analyses are available to summarize the results of hundreds of studies comparing the accuracy of clinical and actuarial decision making in predicting outcomes that range from illness diagnosis and prognosis to future violence and other criminal behavior (Ægisdóttir et al., 2006, Grove et al., 2000; for crime-specific reviews, see Andrews et al., 2006, Hanson & Morton-Bourgon, 2009). In a typical study, trained clinicians synthesize data on a client (from interviews, tests, files, etc.) and then predict an outcome, like violence. Their accuracy is then compared to that of an actuarial prediction in which the same information is used in a formula previously developed based on empirical relations between the predictors and outcome.

The results of these meta-analyses are remarkably consistent with Meehl's (1954) controversial determination that actuarial methods perform as well as, or better than, clinical methods. Based on 41 studies on a range of outcomes, Ægisdóttir et al. (2006) found modest but reliable superiority of the actuarial method over the clinical method (d=-.12). For predicting violence or other criminal behavior specifically, they concluded the actuarial approach was "clearly superior to the clinical approach" (d=-.17).

Importantly, training and experience do little to change this bottom line, countering objections that "those studies of clinical judgment didn't include my professional judgment" (Spengler, 2013). Ægisdóttir et al. (2006) found that even the subset of "best" professionals designated as experts could not outperform statistical formulae. Of course, RAIs vary significantly in quality. But in general, judgments based on them are superior to clinical judgment.

Algorithms typically outperform criminal justice professionals in assessing risk

Most meta-analyses involved mental health professionals, who often serve as experts in justice settings, rather than justice professionals. How accurate are judges' unaided risk assessments? Broadly, the little evidence available indicates judges' typical decision-making processes are much like those of other people—largely intuitive, heuristic-based, and subject to bias (Guthrie, Rachlinski & Wistrich, 2007; Rachlinski, Johnson, Wistrich & Guthrie, 2009). In a rare study, Gottfredson (1999) used a historic sample of 962 felony offenders assessed at sentencing to compare the accuracy of judges' subjective predictions of reoffending with that of predictions made by an actuarial formula (that was not cross-validated). Controlling for time that offenders were at risk for recidivism, the actuarial formula (d= .90) predicted recidivism much more strongly than judges' ratings (d= .54).

These results are echoed by recent comparisons of the accuracy of algorithmic decisions versus judges' decisions about whether to release defendants before trial (Jung, Concannon, Shroff, Goel & Goldstein, 2017; Kleinberg, Lakkaraju, Leskovec, Ludwig & Mullainathan, 2017). Judges' pretrial release decisions were used to approximate risk judgments, since decisions ostensibly turn on prediction of antisocial behavior such as failure to appear. Improving upon past research, these studies addressed a common counterfactual estimation problem, i.e., how one determines the likelihood of recidivism if the algorithm would have released a defendant detained by a judge. The studies answered this question by using causal inference techniques that leveraged the randomness of judges' decisions and the weak relationship between these decisions and actual risk. Using data on nearly 800,000 arrestees subject to pretrial release decisions, Kleinberg et al. (2017) found that replacing judicial decisions with algorithmic decisions could reduce pretrial crime by 25% with no change in the incarceration rate or, alternatively, could reduce jailing rates by 40% without increasing pretrial crime rates. Jung et al. (2017) similarly found that machine-learned decisions outperformed judges—and demonstrated that simple statistically derived rubrics (i.e., the weighted sum of two variables) performed on par with complex algorithms.

Additional evidence that RAIs outperform criminal justice professionals' predictions of recidivism comes from research looking at situations where professionals "override" or adjust an actuarial risk level. Theoretically, justice professionals will beat the actuarial method when they recognize features that rarely occur and countervail the actuarial prediction. Meehl's (1954) classic example is an individual classified in a group with an 80% likelihood of going to the movies tomorrow—except she badly broke her leg today and is immobilized in a hip cast. A more relevant example is an individual classified in a group with a 20% likelihood of proximate recidivism who is expressing specific homicidal intent and has the access and means to carry out this act. However, studies of judges (Krauss, 2004), probation officers, and other correctional professionals (e.g., Cohen, Pendergast & VanBenschoten, 2016; Hanson & Morton-Bourgon, 2009; Wormith, Hogg & Guzzo, 2012) indicate that professional overrides *decrease* accuracy in predicting reoffending, compared to unadjusted actuarial estimates. For example, based on a sample of 3,646 offenders, Guay & Parent (2018) found that probation officers overrode the risk classifications of a commonly

used RAI in 7% of cases—and the unadjusted actuarial estimate predicted new arrests more strongly than the officer's adjusted estimate (d= .87 & .56, respectively).

Structuring professional judgment increases predictive accuracy

As the above example suggests, not all RAIs are fully "actuarial." Skeem and Monahan (2011) explain that RAIs can be arrayed on a continuum of rule-based structure, with completely unstructured ("clinical") assessment occupying one pole of the continuum, completely structured ("actuarial") assessment occupying the other pole, and forms of partially structured assessment lying between the two. Fully actuarial RAIs—like the Virginia instrument (Farrar-Owens, 2013)—structure all four processes of risk assessment, in that they (1) identify risk factors that are empirically valid (and legally acceptable), (2) determine a method for measuring ("scoring") these risk factors, (3) specify a procedure for combining risk factors (e.g., summing scores), and (4) produce the final estimate of risk (e.g., "moderate risk"; "belongs to a group with a 47% recidivism rate"). Partly actuarial RAIs like the Level of Services Inventory (LSI; see Wormith et al., 2012) and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS; see Brennan, Dieterich & Ehret, 2008) structure three processes of risk assessment (identification, measurement, and combination of risk factors), but allow professional judgment to shape the final risk estimate by permitting clinical "overrides" (see above). "Structured professional judgment" (SPJ) instruments like the HCR-20 (see Guy, Kusaj, Packer & Douglas, 2015) structure two processes of risk assessment, specifying a list of risk factors to score on a three-point scale but leaving professionals to rely on their own judgment to combine scores (step 3) and to estimate whether an offender is low, medium or high risk (step 4).

The essential point is that all three types of RAIs—structured professional judgement, partly actuarial RAIs, and fully actuarial RAIs—outperform unaided clinical judgment, but evidence is mixed on whether fully actuarial RAIs outperform the other two. On one hand, professional overrides compromise predictive accuracy, suggesting that fully actuarial approaches are superior. On the other hand, meta-analyses indicate that one well-validated RAI predicts offending as well as another—whether it is fully actuarial or merely structures judgment (e.g., Campbell, French & Gendreau, 2009; Olver, Stockdale & Wormith, 2009). In a meta-analysis of 28 studies that controlled for investigator allegiance and methodological variance, Yang, Wong, and Coid (2010) found the efficiencies of nine RAIs in predicting violence were essentially "interchangeable," with accuracy estimates falling in a narrow band (AUCs= .65-.71). Most studies used summed scores for SPJ instruments (making them more actuarial), but Chevalier's (2017) meta-analysis indicates no significant difference in predictive accuracy between summed scores and professional judgments (low/medium/high risk) on these RAIs.

The Dressel and Farid Study

At first blush, a recent study published in Science Advances seems at odds with the evidence reviewed above. Dressel and Farid (2018) claim to show that laypeople predict recidivism as accurately as a widely used actuarial RAI, the COMPAS. Based on human predictions and COMPAS

predictions for 1,000 defendants, the authors found similar average predictive accuracies (62% for human vs. 65% for COMPAS, $p < .05$).

But a closer look at this study's human predictions suggests the results echo past findings that structured judgment can perform as well as actuarial approaches. Laypeople's judgments were operationalized in a way that constrained inputs, reduced inconsistency, promoted learning and motivation and, as a result, ostensibly lifted accuracy rates. Laypeople were recruited through Amazon's Mechanical Turk to participate in an experiment on "predicting crime" in exchange for money: $1 for completing the task and a $5 bonus for accuracy (i.e., participants received the bonus if their accuracy exceeded 65%, the accuracy of COMPAS). Each participant was shown 50 mini-vignettes that listed a few features of a real defendant in narrative form, i.e., sex, age, current charge, and number of prior adult and juvenile offenses. After each mini-vignette, laypeople indicated whether they thought this person would commit another crime within two years and were instantly informed whether their answer was correct (and their cumulative accuracy) before moving onto the next mini-vignette. Across these responses, the overall accuracy was 62%, comparable to the accuracy of the COMPAS predictions (65%).

This estimate, however, does not characterize the accuracy of *unaided* human prediction. Instead, it indicates what humans can achieve when…

(a) …the only inputs are risk-relevant and consistent across cases. Participants were provided with a few sentences per case that listed robust risk factors for recidivism. This mimics structured checklists that professionals are advised to use to increase consistency and accuracy when making predictions (Guthrie et al., 2007). Even if a professional uses such a checklist, their inputs in real settings involve more thorough and more inconsistent information (e.g., presentence investigation reports, defendants' demeanor, victim impact statements)—much of which is risk-irrelevant or biasing.

(b) …many prediction events are experienced sequentially, interspersed with immediate feedback on accuracy. This created a "kind" environment—i.e., one shown to be ideal for humans to intuitively learn the probabilities of specific outcomes, even when the rules are not transparent (Hogarth & Soyer, 2011). Kind environments promote accuracy, unlike the necessarily "wicked" learning environments that characterize justice settings, where outcomes cannot be observed immediately and are never observed for all cases (Hogarth, Legjarraga & Soyer, 2015; Guthrie et al., 2007).

(c) …a clearly specified prediction goal and incentives to meet that goal are provided. Unlike everyday justice contexts, participants were told to aim for a 65% accuracy rate and provided bonus money for reaching this goal. These are classic, well-validated principles of motivation and behavior change.

These boosted "human predictions" are far removed from unaided human judgment—and, for that matter, from structured professional judgment. In more ecologically valid studies—including field

experiments in pretrial settings (Danner, VanNostrand, & Spruance, 2015; Goldkamp & Gottfredson, 1985)—well-validated algorithms that structure or replace judgment outperform unaided judgment in predicting recidivism.

How structure promotes accuracy—and shows promise in real world justice settings

Using an RAI to structure or replace professional judgment increases predictive accuracy partly because it reduces the noise inherent to human decision making. For example, some judges predict recidivism better than others (Gottfredson, 1999)—and judge-based differences in a defendant's likelihood of pretrial release are large (Jung et al., 2017; Kleinberg et al., 2017). Given the same set of information, two people often disagree about risk. Given the same set of information at two time points, the same person can arrive at different risk estimates. When the risk assessment process is fully structured, actuarial RAIs assign optimal weights to variables and consistently apply well-specified rules to yield reproducible results. Given the same inputs, these RAIs generate the same risk estimate each time—they do not have off days.

Whether risk estimates are more accurate when they structure or replace professional judgment is arguably a moot point. In justice settings, algorithms and professional judgment must work together to promote accuracy because risk estimates rarely provide dispositive answers to legal questions. There is preliminary evidence that professionals can implement RAIs effectively in their efforts to reduce incarceration without compromising public safety (Danner et al., 2015). For example, in an experiment conducted in Philadelphia (Barnes et al., 2010), the Adult Probation and Parole Department used an RAI to identify community supervisees at low risk of violence and decreased their supervision levels without increasing crime rates.

PART II: EQUITY

Actuarial risk assessment instruments work by identifying statistical patterns in historical records. For example, one might start with information on the attributes of past defendants (e.g., their age and number of prior arrests), judicial decisions (e.g., release or detain), and outcomes (e.g., whether the individual engaged in future criminal activity). A statistical model is then constructed to estimate the empirical likelihood that released defendants in the historical data reoffended. Assuming future defendants are similar to those in the data, the constructed model can then be used to predict the behavior—and hence gauge the risk—of previously unseen individuals.

In 2016, a widely read investigative news story alleged that one such actuarial RAI, the COMPAS, was "biased against blacks" (Angwin et al., 2016). Prompted in large part by that article, researchers and practitioners have since voiced deep concerns that statistical risk assessments might inadvertently discriminate, particularly against groups defined by race and gender. We enumerate and examine several of those concerns, starting with potential problems in the data that can, if not addressed, exacerbate historical inequities. We then introduce—and note the limitations of—several popular mathematical measures of fairness that have been proposed to detect and

mitigate such bias. We conclude this discussion by offering advice for constructing equitable risk assessment tools.

In the end, as with all tools, one must consider the value of imperfect RAIs relative to the available alternatives—most commonly, unaided human judgment which, as we discussed above, is susceptible to its own inaccuracies and biases.

Bias in the data

A common set of misgivings about RAIs revolves around the historical data used in their construction—called "training data". Many have expressed skepticism that risk assessments can ever be fair, as the training data necessarily contain inaccuracies, some of which arise through biases in past human actions. The two main concerns can be summarized as: (1) *measurement error*, the discrepancy between reality and its representation in the data; and (2) *sample bias*, the discrepancy between the training data and the population of individuals to which a constructed model is ultimately applied. We discuss both issues in turn below.

Measurement error. In order to estimate, for example, the risk that a defendant would commit a crime if released before trial, it is important that both the attributes used to make the prediction and the outcome being predicted are measured accurately. Mismeasurement in the attributes is commonly termed *feature bias*, whereas mismeasurement in the outcomes is called *label bias*. It is often possible to statistically account for feature bias, but it is considerably harder to deal with label bias. Indeed, this latter issue is arguably one of the most serious facing the design of equitable risk assessment tools.

We illustrate feature bias with a simple example. Suppose that one's likelihood of future criminal activity increases with the number of past drug sales one has carried out. Since the actual number of drug sales an individual has engaged in is unlikely to be recorded, it is common to use the number of past *arrests* for drug sales as a proxy. However, minorities who engage in drug-related crime are more likely to be arrested than whites who engage in the same behavior (Ramchand et al., 2006). As a result, using recorded drug arrests as a proxy for actual drug sales may (incorrectly) rate black defendants as higher risk than white defendants who have engaged in similar criminal behavior.

One potential solution to this feature bias problem is to fit two separate statistical models, one for black defendants and another for white defendants. In the absence of label bias (i.e., if outcomes, like recidivism, are accurately measured), this strategy would result in a model that correctly discounts for the longer criminal histories of black defendants. For example, such a model might discover that a black defendant with two drug arrests is about as risky as a white defendant with one drug arrest. In practice, it can be legally challenging, though not impossible, to base risk assessments on race or gender, a point we elaborate on below. But from a purely statistical point of view, the problem of feature bias can often be overcome.

In contrast to feature bias, label bias presents a conceptually similar though much harder problem to counter. Suppose one estimates the likelihood a defendant *commits* a new crime (the outcome, which is hard to observe) by instead estimating the likelihood a defendant is *convicted* of a new crime (a proxy which is often readily available). As above, high-intensity policing in certain neighborhoods may result in minorities being arrested and convicted more often than whites who commit the same offenses (Lum and Isaac, 2016), and as a result, the mismatch between outcomes and proxies may lead one to systematically overestimate the risk posed by black defendants relative to white defendants.

Unlike the analogous problem with feature bias, fitting separate statistical models does not help when the outcome measure itself is corrupted. In general, there is unfortunately no perfect solution to label bias. In practice, however, one might focus on predicting outcomes (such as *violent* crime) that are believed to be more accurately recorded, or at least where the available proxies are less racially skewed (Skeem and Lowenkamp, 2016).

Sample bias. When algorithms are trained on data that do not reflect the population to which they are applied, potentially discriminatory consequences can result. One recent study found that commercial facial recognition software designed to infer gender performed worse on dark-skinned individuals compared to light-skinned individuals (Buolamwini and Gebru, 2018). These differences in accuracy are likely due in part to the lack of dark-skinned faces in widely used facial recognition datasets.

In the context of criminal justice risk assessment, it can be logistically challenging to develop instruments that are customized for local populations. For example, the popular Ohio Risk Assessment System (ORAS) was developed on a sample of several hundred defendants in Ohio but is now used nationwide (Latessa et al., 2010). If defendants in other jurisdictions differ systematically from those in Ohio, the ORAS risk assessments could yield inaccurate estimates. More recently developed tools attempt to mitigate this issue by using training data from counties across the country (Milgram et al., 2014). While an important step forward, this approach is not a complete solution, as a model trained on national data may still not account for the idiosyncrasies of every jurisdiction. Unfortunately, it is often infeasible to develop truly localized models, particularly in smaller jurisdictions that lack adequate historical data to train models that perform well.

Formal definitions of fairness and their limitations

In part due to concerns with the training data, researchers have increasingly sought out metrics both to gauge the fairness of existing risk assessment tools and to design more equitable ones. In particular, three broad classes of fairness definitions have gained prominence in the academic community. The first, which we call *anti-classification*, requires that risk assessment algorithms not

consider protected characteristics—like race, gender, or their proxies—when deriving estimates.[1] The second class of definitions demand *classification parity*, meaning that certain common measures of predictive performance (like false positive or negative rates) be equal across groups defined by the protected attributes. For example, one might require that among defendants who do not go on to reoffend, an equal proportion of white and black defendants are classified by the algorithm as high risk—a criterion that ensures false positive rates are equal. Finally, the third formal fairness definition, known as *calibration*, requires that outcomes are independent of protected attributes after controlling for estimated risk. For example, among defendants estimated to have a 10% chance of reoffending, calibration requires that whites and blacks indeed reoffend at similar rates.

These formalizations of fairness each have intuitive appeal. It can feel natural to exclude protected characteristics in a drive for equity. Likewise, one might understandably interpret differences in error rates as indicating problems with the algorithm's design (e.g., sample bias in the data on which it was trained), or as promoting social injustices. However, perhaps surprisingly, all three of these popular definitions of algorithmic fairness—anti-classification, classification parity, and calibration—suffer from deep statistical limitations. In particular, they are poor measures for detecting discriminatory algorithms and even more importantly, designing algorithms to satisfy these definitions can, perversely, negatively impact the well-being of minority and majority communities alike (Corbett-Davies and Goel, 2018; Mayson, 2019). We briefly discuss the limitations of each of these measures in turn below.

Anti-classification. In some cases, it may be necessary for risk assessment algorithms to explicitly consider protected characteristics to achieve equitable outcomes. As discussed above, one can in theory combat feature bias by using group-specific risk assessments. To give another example, we note that women are typically less likely to commit a future crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. For example, in Broward County, Florida, women with a COMPAS risk score of 6 out of 10 (making them "medium risk") reoffend at about the same rate as men with a risk score of 4 (making them "low risk") (Corbett-Davies and Goel, 2018).[2]

Recognizing this problem, some jurisdictions have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women. Though some might consider this choice controversial, the Wisconsin State Supreme court affirmed the application of such gender-specific

[1]The term "anti-classification" is popular among legal scholars, but it is not commonly used by computer scientists or statisticians working in this field. In general, given the interdisciplinarity and nascency of fair machine learning, a variety of terms are often used by different authors to describe the same underlying concept.

[2]The fact that men and women with similar criminal histories recidivate at different rates is not necessarily due to inaccuracies in recorded data; it may simply be the case that the relationship between predictive attributes and recidivism differs by gender.

tools, writing that "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose" (Wisconsin v. Loomis, 2016).

Classification parity. In Broward County, the false positive rate for COMPAS risk assessments is twice as large for black defendants than white defendants. Specifically, among those who did not go on to reoffend, 31% of black defendants were rated medium or high risk, compared to 15% of white defendants. This stark difference was the basis for Angwin et al.'s (2016) assertion that the COMPAS algorithm is racially biased. Accordingly, some have called for risk assessment tools to ensure such error rates are equal across groups.

Counterintuitively, though, differences in false positive rates often tell us more about the underlying populations than about bias in the algorithm (Corbett-Davies and Goel, 2018). False positive rates can mechanically increase with a group's overall rate of recidivism. In Broward county, black defendants appear to reoffend more often than whites, and so a higher false positive rate is an expected consequence of any algorithm that accurately captures each individual's risk. This pattern would similarly hold even if risk assessments were based on unaided human judgment rather than a statistical model.

This general statistical phenomenon affects nearly every commonly used measure of accuracy. (It is called the problem of *infra-marginality*, and has been discussed, for example, by Ayres, 2002; Corbett-Davies and Goel, 2018; and Simoiu et al., 2017). As a result, examining between-group differences in error rates is a poor means for assessing fairness.

Further, demanding error rates be equal can itself lead to discriminatory decision making. Achieving such parity often requires implicitly or explicitly misclassifying low-risk members of one group as high-risk, and high-risk members of another as low-risk, potentially harming members of all groups in the process. For example, to equalize false positive rates in the Broward County COMPAS data, one could classify black defendants as risky if they score a 6 or higher, but classify white defendants as risky if they score a 4 or higher (Corbett-Davies et al., 2017). This double standard raises clear concerns about equity, and illustrates the problem with using classification parity as a fairness metric.

Calibration. Finally, we turn to calibration. For a tool to be calibrated, defendants with similar scores must in reality reoffend at similar rates, regardless of group membership. For example, among those with a risk score of 8, approximately the same fraction of white defendants and black defendants should reoffend if released.

Calibration is generally a desirable property for a risk assessment instrument to have, and indeed many of the most popular tools are calibrated across race groups. Perhaps surprisingly, however, it provides only a weak guarantee of equity. The illegal practice of redlining in banking illustrates how one can strategically discriminate while maintaining calibration (Corbett-Davies and Goel, 2018). To unfairly limit loans to minority applicants, a bank could base risk estimates only on coarse information, like one's neighborhood, and ignore individual-level factors, like income and credit

history. The resulting risk scores would be calibrated—assuming majority and minority applicants default at similar rates within neighborhood—but could be used to deny loans to creditworthy minorities who live in relatively high-risk neighborhoods.

While such strategic discrimination may be less common today, similar effects can arise from inexperience rather than malice. For example, algorithm designers may inadvertently neglect to include important predictors in risk models, resulting in risk scores that are insufficiently individualized.

A pragmatic view of fairness. In contrast to the metrics described above, practitioners have long designed tools that adhere to an alternative fairness concept. Namely, after constructing risk scores that best capture individual-level risk—and potentially including protected traits to do so—similarly risky individuals are treated similarly, regardless of group membership. The operative word here is "best", and one must carefully consider all the available information to create the most accurate risk assessments. Selectively excluding information can lead to the type of discriminatory redlining effects mentioned above.

Using this notion of fairness, decision makers, when determining which defendants to release while they await trial, could first select an acceptable risk level and then release those individuals estimated to fall below that threshold. This policy follows widely accepted legal standards of equity. Further, such a decision strategy—with an appropriately chosen decision threshold—maximizes a natural notion of social welfare for all groups. Importantly, however, this thresholding approach will in general violate classification parity, and may additionally violate anti-classification, as producing accurate risk assessments might require using protected characteristics. The underlying risk scores will typically satisfy calibration, but the goal is to do so by providing accurate individual-level predictions that avoid inequities of the type illustrated in our redlining example.

Designing equitable algorithms

How, then, can one design equitable algorithms? Our discussion above of measurement error and sample bias immediately implies several design principles for mitigating these issues. First, if it promotes accuracy—and if legally permissible—consider fitting separate models by gender or other protected group characteristics. This explicit consideration of group membership can mitigate feature bias by accounting for relationships between risk factors and outcomes that may differ between groups. Second, when possible, predict outcomes that are accurately measured, like arrests for violence but not for drug crimes, to avoid label bias—and work to improve data collection procedures as necessary to do so. Third, to mitigate sample bias, train risk assessments on data collected from jurisdictions where they are intended to be applied.

We further caution against forcing equal false positive rates across protected groups to achieve classification parity, as such an approach can itself harm majority and minority groups alike. Equalizing false positive rates necessarily means misclassifying individuals, leading to relatively low-risk members of one group detained and relatively high-risk members of another released.

Accordingly, one group faces unnecessary incarceration while another bears the burden of high-risk individuals being released into the community.

We conclude by making three high-level recommendations for addressing fairness concerns in the context of algorithmic risk assessments. First, both technical and policy discussions of fairness should be grounded in real world costs and benefits, such as potential effects on public safety and on the number of individuals incarcerated. While it is often unclear exactly how to quantify costs and benefits of algorithmic interventions, strict adherence to formal mathematical conceptions of fairness addresses these issues indirectly, if at all.

Second, the task of estimating risk of reoffending should not be conflated with the task of intervening, based on estimated risk, to prevent reoffending. An algorithm may estimate that a particular defendant, if released, has a high risk of failing to appear in court—but this risk estimate does not automatically translate to real world action, like financial assistance, enhanced supervision, or detention. The goal of a risk assessment instrument should be to estimate *risk* as accurately as possible. On the basis of such estimates, policymakers should then determine whether the costs and benefits of particular interventions achieve society's goals. For instance, one may decide that the costs of detaining an individual who is the primary financial provider for their family may be higher than the costs of detaining an individual with no dependents, and apply different interventions, even if the individuals are similarly risky.

Finally, we encourage transparency, both in the development and the application of risk assessment tools. Transparency helps ensure that risk models are designed with the best available statistical methods and training data, promoting accuracy. Transparency further builds confidence in risk assessment instruments by helping judges, defendants, community members, and other stakeholders understand and evaluate these tools.

PART III: JURISPRUDENCE

The use of algorithms in the criminal justice system clearly raises important issues. Unfortunately, legal decision makers—whether one looks at legislatures or courts—have either ignored these issues or only reluctantly and half-heartedly addressed them. Much more attention to the jurisprudence of algorithm-aided decision making is necessary.

The law regarding the accuracy and relevance of prediction evidence

Legislatures and courts have long taken a casual approach to risk assessment. The most glaring example is the Supreme Court's decision in *Barefoot v. Estelle* (1983), which held that the introduction of concededly highly unscientific testimony about dangerousness does not violate the Constitution, even when proffered by the state in a capital sentencing proceeding. Unconstrained by formal rules of evidence in either pretrial or post-conviction settings, courts allow virtually any type of submission about risk, whether it comes from probation officers or mental health professionals, and whether framed in actuarial or clinical terms. Judicial rejection of challenges to

prediction testimony often merely state that such testimony is necessary to achieve the state's ends, with very little analysis of the accuracy or methodology of the expert (Faigman et al., 2018).

This judicial nonchalance should change. Pretrial detention and enhanced sentences should not be based on a risk assessment unless it meets basic indicia of reliability. While *Barefoot* held that the Constitution's due process clause does not mandate such a requirement, the Supreme Court's holdings in *Daubert v. Merrell Dow Pharmaceuticals* (1993) and its progeny (see, e.g., *General Electric v. Joiner*, 1997), now followed in a majority of states, make clear that the statutory rules of evidence applicable at trial require judges to evaluate the scientific value of expert testimony. Given the deprivation of liberty at stake, *Daubert* should apply to pretrial and post-conviction proceedings as well. If *Daubert* and the rules of evidence governed the use of risk assessment instruments, judges would have to assess whether the instrument in question is "reliable," including, according to *Daubert*, whether it has been subject to scientific testing on a population similar to the offender's, whether its error rates are available, whether it has been subject to peer review, and whether it is generally accepted in the field of prediction (*Daubert*, 1993, 593-94). In other words, the types of considerations canvassed in previous sections of this chapter would need to be addressed.

Just as important from a legal perspective is *Daubert's* additional injunction that the expert evidence in question "fit" the legal proposition at issue. As *Daubert* stated, "'Fit' is not always obvious, and scientific validity for one purpose is not necessarily scientific validity for other, unrelated purposes" (591). In the risk assessment context, the fit issue has been almost completely ignored by the courts. Even the instrument that most accurately predicts reoffending may not be a good legal fit if it does not help answer the specific questions the law wants answered.

Presumably, courts making pretrial and sentencing decisions would want information about four issues: (1) the probability *P*, (2) that behavior *Y*, (3) will occur during period of time *T*, (4) if intervention *Z* is taken (see Slobogin, 2018). The probability question requires determining how the legal standard of proof (e.g., beyond a reasonable doubt, clear and convincing evidence) interacts with the legal definition of risk (which could be equated with a risk estimate, e.g., a 10%, 20% or 30% likelihood of recidivism). The outcome question requires determining the type of antisocial conduct (e.g., against person; felony v. misdemeanor; arrest v. conviction) that, if predicted with the requisite probability, justifies intervention in the legal context in which the prediction takes place. The timing question requires consideration of how long the intervention may be imposed (e.g., a few months, several years) before another evaluation is necessary. And the intervention question requires determining what type of action (e.g., detention; restrictions on travel; treatment; electronic monitoring) is necessary to prevent the predicted harm.

Unfortunately, in many jurisdictions neither the relevant statutes nor the interpretive caselaw answer any of these questions. In the pretrial setting, the federal statute requires "clear and convincing evidence" that no condition other than detention "will reasonably assure the safety" of others (Federal Bail Reform Act, 1984). But the courts have not specified in quantifiable terms what qualifies as "clear and convincing evidence" or what constitutes "reasonable" assurance of safety. At sentencing the relevant provisions are similarly vague. Some statutes that permit or mandate

diversion for offenders who are "low risk" merely state that proviso, with no further explanation of what low risk means and with no standard of proof indicated (see, e.g., the description of Virginia's sentencing law in Kern & Farrar-Owens, 2004). The situation is not much better at capital sentencing, despite the supposed enhanced concern about due process in that context. For instance, in Texas, where the death penalty statute requires proof beyond a reasonable doubt that an offender "will commit criminal acts of violence that constitute a continuing threat to society," courts have held that "the Legislature declined to specify a particular level of risk or probability of violence," and thus have left the decision about risk to the complete discretion of the judge or jury (*Coble v. State*, 2010).

The law regarding the fairness of predictive algorithms

One likely reason for this stunning judicial abdication is that, until recently, prediction testimony was itself extremely vague. Perhaps the courts have felt that they could not demand answers to questions that could not be answered in other than a subjective way. But the latter difficulty has diminished with the advent of evidence-based risk assessment. As the above discussion indicates, such instruments can provide relatively precise probability estimates of violence or general recidivism, within designated time periods. They may also identify, in a more structured way than was previously the case, changeable factors that theoretically would reduce risk if targeted with appropriate treatment—although evidence that these changeable factors are causally related to recidivism is in short supply (see Skeem et al., 2017). The point is that the courts can and should demand data-based answers to questions about risk assessment and risk reduction.

A few courts have done so, but their attempts fall short in a number of ways. The leading case to date in this regard is the aforementioned Wisconsin Supreme Court decision in *Wisconsin v. Loomis* (2016), which involved a challenge to the COMPAS, a relatively complex risk assessment tool that was used to assess Loomis' risk and that was, in part, the basis for the sentence he received. Loomis argued that his sentence violated due process in several respects. First, he argued that because the company that developed the COMPAS, Equivant (formerly Northpointe), would not release the code underlying the instrument's algorithm, he was prevented from analyzing its accuracy. Second, he contended that, because his risk score was based on data derived from a group, his sentence was not "individualized." Third, he argued that, because the COMPAS includes male sex as a risk factor, it discriminated on the basis of gender.

The Wisconsin Supreme Court rejected the first argument on the ground that Loomis had access to the instrument itself and could roughly determine how his risk score was produced based on the answers he and public records provided. At the same time, the court made a bow to Loomis' concerns by requiring that, henceforth, trial courts must be informed of Equivant's trade secret claim, as well as of the facts that the COMPAS had not been normed on a Wisconsin population, that it might be biased against minorities (this requirement relied upon the faulty logic about false positive rates discussed in Part II), and that it should be periodically re-validated. With respect to the second, failure-to-individualize, argument, the court agreed that COMPAS scores are only able to identify "groups of high-risk offenders—not a particular high-risk offender," and mandated that

lower courts be made aware of that fact as well (265). But ultimately it also refused to reverse Loomis' sentence on this ground, because results such as those provided by the COMPAS can be "helpful" to sentencing courts and should be consulted despite their generalized nature as long as they are not dispositive of the risk determination. Finally, on the discrimination issue, the court pointed out that excluding gender from the COMPAS, as Loomis requested, would make the risk score less accurate and tend to overestimate the risk that females posed. While the court thus refused to overturn Loomis' sentence, it ended by cautioning that "using a risk assessment tool to determine the length and severity of a sentence is a poor fit," and thus should only be used for such matters as (1) "diverting low-risk prison-bound offenders to a non-prison alternative; (2) assessing whether an offender can be supervised safely and effectively in the community; and (3) imposing terms and conditions of probation, supervision, and responses to violations." It also repeated that in no case should the risk score be "determinative" (272).

The *Loomis* court is to be commended for its willingness to address difficult issues connected with risk assessment. But its reasoning is flawed in several respects. First, for reasons that should be clear from previous sections of this chapter, without transparency neither the offender nor the court can assess which risk factors were included, what weights were assigned to them in estimating risk, and a variety of other important scientific matters. Thus, the court's willingness to honor Equivant's trade secret claim is unfortunate. In *Gardner v. Florida* (1977), the U.S. Supreme Court held that persons subject to sentence (at least a capital sentence) are entitled to know about and test the accuracy of the information heard by the sentencing authority. That ruling should require private companies to provide criminal defendants and courts with the information needed to evaluate accuracy. Concerns about giving competitors an advantage or discouraging innovation are overblown, especially if protective orders or *in camera* review requirements are imposed; further, subjecting risk algorithms to the adversarial process is likely to improve rather than undermine their quality (see Wexler, 2018).

The court was correct to discount Loomis' concern about the lack of individualization in his sentence. But its rationale for doing so—its admonition that risk assessment scores should be only one of the factors considered by the court in determining risk—is problematic. Of course, offenders should always be able to introduce evidence of protective factors that were not considered in the development of the state's instrument, such as treatment successes, recent changes in circumstances, or aspects of criminal history—like a wrongful arrest—that undercut the factual basis for the risk score. But telling judges they can substitute their own assessment for a risk score does not make sense from a scientific point of view to the extent the variables the judge considers were already explicitly tested in constructing the instrument; just as importantly, as illustrated by Part I's discussion of how professional 'overrides' of actuarial estimates can backfire, it could well reintroduce the bias that instruments are designed to prevent. Moreover, the court's acceptance of the premise of Loomis' argument—that risk assessment instruments are suspect because based on group data—is off-base. While risk instruments are derived from offenders other than the examinee, all expert testimony—including non-actuarial prediction testimony—is ultimately based on assumptions about the kind of person an offender is, as is the judge's ultimate determination of

risk (Faigman et al., 2014); the key difference, and one that should count as an advantage, is that the instrument displays its stereotyping assumptions on its face.

Third, while the court is correct about the effect on accuracy of removing variables like gender from a risk instrument, its reasoning glosses over two fundamental concerns underlying Loomis' final objection. The first is an equal protection argument, to the effect that such instruments, on their face, discriminate on the basis of gender. In fairness to the court, Loomis did not directly raise an equal protection claim. But such a claim was implicit in his due process argument. Thus, the court's response to the effect that the failure to consider gender would inaccurately assign women higher risk scores, while true, should have been augmented with an analysis of why the state's interest in avoiding such inaccuracy is compelling enough to overcome the use of an instrument that explicitly relies on gender to reach its conclusions (cf. *United States v. Virginia*, 1996).

The more important concern raised by Loomis' third objection (albeit again one that that Loomis did not directly raise) is that a sentence grounded even in part on gender could be considered antithetical to the idea that punishment should be based on blameworthy conduct. In *Buck v. Davis* (2017), the U.S. Supreme Court stated:

> It would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race. . . .[That would be] a disturbing departure from a basic premise of our criminal justice system: *Our law punishes people for what they do, not who they are*" (778).

The italicized language suggests that not only race and gender, but age, diagnosis and any other risk factors that are not based on (blameworthy) conduct are illegitimate grounds for punishment. Taken literally, *Buck's* restriction could severely degrade the accuracy of risk assessment instruments, both at sentencing and in connection with pretrial detention, to the extent such detention is seen as a form of punishment (cf. *Bell v. Wolfish*, 1979). Read most expansively, the language could even call into question whether risk is ever a legitimate concern in the criminal justice system, as risk necessarily associates punishment with anticipated future acts, not only what a person has done (see Slobogin, forthcoming).

However, the Supreme Court probably does not mean its statement in *Buck* to be taken literally. On several occasions it has even upheld death sentences imposed after a finding of dangerousness based in part on diagnosis (see, e.g., *Barefoot*, 1983). At bottom, *Buck* appears to be a case about race, not about all immutable traits or risk more generally.

Assuming that risk continues to play a prominent role in pretrial and sentencing decision making, a final problem with the *Loomis* decision is that it left completely open the aforementioned fit issues that should be addressed in assessing risk (concerning the requisite probability, outcome, timing and intervention options). Perhaps the court is right that the COMPAS is a "poor fit" for determining the length and severity of a sentence. But, if so, the instrument should not be used even to determine whether a person can be diverted from prison, since those who are *not* diverted because

of their COMPAS score are in effect having the severity, if not the length, of their sentence determined by it. Mandating, as *Loomis* does, that the COMPAS not be "determinative" of one's risk or of one's sentence disingenuously avoids the issue. As an Iowa court subsequently put it, "We are not persuaded that the difference between *reliance upon* and *consideration of* these actuarial estimates saves the sentencing process" (*Iowa v. Gordon*, 2018, emphasis in original). (And, in any event, as noted above, in many cases reliance on these instruments may be scientifically preferred on the issue of risk). It would have been better if the *Loomis* court had straightforwardly stated that risk assessment instruments may be used to assess risk, but only if they provide information that helps answer the relevant legal questions. The court should then have identified precisely what it thought those questions should be.

CONCLUSION

Well-designed predictive algorithms can provide information about defendant and offender risk that is more accurate and less biased than clinical decision making. But the full potential of risk assessment instruments can only be realized if the algorithms are properly constructed and properly applied by the legal system. This chapter has outlined the scientific and legal challenges to achieving those goals.

REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. The Counseling Psychologist, 34(3), 341-382.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. Crime & Delinquency, 52(1), 7-27.

Angwin, Julia, Jeff Larson, Surya Mattu & Lauren Kirchner, Machine Bias, ProPublica, May 23, 2016.

Ayres, Ian, Outcome tests of racial disparities in police practices, Justice Research and Policy, 131-142 (2002).

Barefoot v. Estelle (1983). 463 U.S. 880.

Barnes, G. C., Ahlman, L., Gill, C., Sherman, L. W., Kurtz, E., & Malvestuto, R. (2010). Low-intensity community supervision for low-risk offenders: a randomized, controlled trial. Journal of Experimental Criminology, 6(2), 159-189.

Bell v. Wolfish (1979). 441 U.S. 420.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. Criminal Justice and Behavior, 36(1), 21–40.

Buck v. Davis (2017). 137 S.Ct. 759.

Buolamwini, Joy and Timnit Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in Proceedings of the Conference on Fairness, Accountability and Transparency (2018).

Campbell, M.A., French, S, & Gendreau, P. (2009). The prediction of violence in adult offenders: a meta-analytic comparison of instruments and methods of assessment. Criminal Justice & Behavior, 36, 567–590.

Chevalier, C. S. (2017). The Association between Structured Professional Judgment Measure Total Scores and Summary Risk Ratings: Implications for Predictive Validity. Unpublished dissertation available at: https://shsu-ir.tdl.org/handle/20.500.11875/2228

Coble v. State (2010). 330 S.W.3d 352 (Tex. Crim. App.).

Cohen, T. H., Pendergast, B., & VanBenschoten, S. W. (2016). Examining overrides of risk classifications for offenders on federal supervision. Federal Probation, 80, 12-21.

Corbett-Davies, Sam and Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (Aug 14, 2018) (unpublished manuscript, available at https://arxiv.org/abs/1808.00023).

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, Algorithmic decision making and the cost of fairness, in Proceedings of the International Conference on Knowledge Discovery and Data Mining (2017).

Danner MJ, VanNostrand, M, & Spruance, LM (2015). Risk-based pretrial release recommendation and supervision guidelines: Exploring the effect of officer recommendations, judicial decision-making, and pretrial outcome. Unpublished report available at: https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/risk-based-pretrial-release-recommendation-and-supervision-guidelines.pdf

Daubert v. Merrell Dow Pharmaceuticals (1993). 509 U.S. 579.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

Faigman, David, Edward K. Cheng, Jennifer L. Mnookin, Erin E. Murphy, Joseph Sanders & Christopher Slobogin (2018). Modern Scientific Evidence: The Law and Science of Expert Testimony (Danvers, Mass: Thomson-Reuters).

Faigman, David, John Monahan & Christopher Slobogin (2014). Group to Individual (G2i) Inference in Scientific Expert Testimony. University of Chicago Law Review, 81, 417-480.

Farrar-Owens, M. (2013). The evolution of sentencing guidelines in Virginia: an example of the importance of standardized and automated felony sentencing data. Federal Sentencing Reporter, 25(3), 168-170.

Federal Bail Reform Act (1984). 18 U.S.C. §§ 3141-3150.

Gardner v. Florida (1977). 430 U.S. 349.

General Electric v. Joiner (1997). 522 U.S. 136.

Goldkamp, J. S., & Gottfredson, M. R. (1985). Policy guidelines for bail: An experiment in court reform. Philadelphia: Temple University Press.

Gottfredson, D. M. (1999). Effects of judges' sentencing decisions on criminal careers. US Department of Justice, Office of Justice Programs, National Institute of Justice. Available at: https://www.ncjrs.gov/pdffiles1/nij/178889.pdf

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12(1), 19-30.

Guay, J. P., & Parent, G. (2018). Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels With the LS/CMI. Criminal Justice and Behavior, 45(1), 82-100.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. Cornell Law Review, 93, 1-44.

Guy LS, Kusaj C, Packer IK, & Douglas KS (2015). Influence of the HCR-20, LS/CMI, and PCL-R on decisions about parole suitability among lifers. Law Hum. Behavior, 39, 232–243.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. Psychological Assessment, 21(1), 1-21.

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus non-transparent description. Journal of Experimental Psychology: General, 140(3), 434-463.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. Current Directions in Psychological Science, 24(5), 379-385.

Iowa v. Gordon (2018). 2018 WL 2084847 (Iowa Court of Appeals).

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. Unpublished manuscript available at: https://arxiv.org/pdf/1702.04690

Kern, Richard P. & Meredith Farrar-Owens (2004). Sentencing Guidelines with Integrated Offender Risk Assessment, Federal Sentencing Reporter, 16: 165-169.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237-293.

Krauss, D. A. (2004). Adjusting risk of recidivism: Do judicial departures worsen or improve recidivism prediction under the Federal Sentencing Guidelines?. Behavioral Sciences & the Law, 22(6), 731-750.

Latessa, Edward, Richard Lemke, Matthew Makarios, Paul Smith & Chris Lowenkamp, The creation and validation of the Ohio Risk Assessment System (ORAS). Fed. Probation (2016).

Lum, Kristian and Isaac, William, To predict and serve?, Significance, 14-19 (2016).

Mayson, Sandra, Bias In, Bias Out, 128 Yale Law Journal (2019).

Meehl, P.E. (1954). Clinical versus statistical prediction. Minneapolis: University of Minnesota Press.

Milgram, Anne, Alexander Holsinger, Marie VanNostrand & Matthew Alsdorf, Pretrial risk assessment: Improving public safety and fairness in pretrial decision making, Federal Sentencing Reporter, (2014).

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. Criminal Justice and Behavior, 36(4), 329-353.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does Unconscious Racial Bias Affect Trial Judges?. Notre Dame Law Review, 84(3), 1195-1246.

Ramchand, R., R. L. Pacula & M. Y. Iguchi, Racial differences in marijuana-users risk of arrest in the United States, Drug and Alcohol Dependence, 264-272 (2006).

Simoiu, Camelia, Sam Corbett-Davies & Sharad Goel, The problem of infra-marginality in outcome tests for discrimination, Annals of Applied Statistics, 1193-1216 (2017).

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. Current Directions in Psychological Science, 20(1), 38-42.

Skeem, J. L., Kennealy, P. J., Tatar, J. R. II, Hernandez, I. R., & Keith, F. A. (2017). How well do juvenile risk assessments measure factors to target in treatment? Examining construct validity. Psychological Assessment, 29(6), 679-691.

Skeem, Jennifer and Christopher Lowenkamp, Risk, race, and recidivism: predictive bias and disparate impact, Criminology, 680-712 (2016).

Slobogin, Christopher (2018). Principles of Risk Assessment: Sentencing and Policing. Ohio State Journal of Criminal Law, 15: 583-596.

Slobogin, Christopher (forthcoming). A Defense of Modern Risk-Based Sentencing, in Risk and Retribution: The Ethics and Consequences of Predictive Sentencing (Jan de Keijser, Jesper Rysberg & Julian Roberts eds.) (London, U.K.: Hart Publishing).

Spengler, P. M. (2013). Clinical versus mechanical prediction. In I. Weiner, J. Graham & J. Naglieri (Eds), Handbook of psychology: Assessment psychology, 26-49. New York: Wiley.

United States v. Virginia (1996). 518 U.S. 515.

Wexler, Rebecca (2018). Life, Liberty and Trade Secrets: Intellectual Property in the Criminal Justice System, Stanford Law Review, 70, 1343-1429.

Wisconsin v. Loomis (2016). 881 N.W.2d 749 (Wisconsin Supreme Court).

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. Criminal Justice and Behavior, 39(12), 1511-1538.

Yang, M., Wong, S. C., & Coid, J. (2010). The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. Psychological Bulletin, 136(5), 740-767.