Risk Assessment Tools in the Criminal Legal System – Theory and Practice:

A RESOURCE GUIDE

A Report Commissioned by the NACDL Task Force on Risk Assessment Tools

NATIONAL
ASSOCIATION
OF CRIMINAL
DEFENSE LAWYERS

NACDL
FOUNDATION FOR
CRIMINAL
JUSTICE

November 2020

Melissa Hamilton

Reader in Law & Criminal Justice
University of Surrey School of Law
J.D., Ph.D., The University of Texas at Austin

For more information contact:

NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS©

1660 L Street NW, 12th Floor, Washington, DC 20036, Phone 202-872-8600

www.NACDL.org

NATIONAL
ASSOCIATION
OF CRIMINAL
DEFENSE LAWYERS

NACDL
FOUNDATION FOR
CRIMINAL
JUSTICE

This publication is available online at

www.NACDL.org/RiskAssessmentReport

Risk Assessment Tools in the Criminal Legal System – Theory and Practice:

A RESOURCE GUIDE

A Report Commissioned by the NACDL Task Force on Risk Assessment Tools

Christopher W. Adams
President, NACDL
Charleston, South Carolina

Lisa M. Wayne
President, NFCJ
Denver, Colorado

Norman L. Reimer
Executive Director, NACDL & NFCJ
Washington, DC

Task Force on Risk Assessment Tools

Vicki H. Young
Co-Chair
Palo Alto, California

Marvin Schechter
Co-Chair
New York, New York

Franklin Draper
Washington, DC

Nina J. Ginsberg
Alexandria, Virginia

Stephen Ross Johnson
Knoxville, Tennessee

Matthew Knecht
New York, New York

John Philipsborn
San Francisco, California

Gail Shifman
San Francisco, California

Colette Tvedt
Denver, Colorado

CeCelia Valentine-Andrews
Vancouver, Washington

AUTHOR
Dr. Melissa Hamilton

NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS

NACDL FOUNDATION FOR CRIMINAL JUSTICE

Table of Contents

I. About NACDL and the NACDL FCJ

The National Association of Criminal Defense Lawyers (NACDL) is the preeminent organization in the United States advancing the mission of the nation's criminal defense bar to ensure justice and due process for persons accused of crime or other misconduct. NACDL envisions a society where all individuals receive fair, rational, and humane treatment within the criminal justice system. NACDL's mission is to serve as a leader, alongside diverse coalitions, in identifying and reforming flaws and inequities in the criminal justice system, and redressing systemic racism, and ensuring that its members and others in the criminal defense bar are fully equipped to serve all accused persons at the highest level.

Founded in 1958 as the professional bar association of the nation's criminal defense attorneys, NACDL's direct membership now includes thousands of direct members in 28 countries, along with 90 state, provincial, and local affiliate organizations totaling approximately 40,000 attorneys, including private criminal defense lawyers, public defenders, active U.S. military defense counsel, law professors, judges and others whose work evinces a commitment to preserve fairness within America's criminal justice system.

The NACDL Foundation for Criminal Justice (NFCJ) is a 501(c)(3) non-profit entity that supports NACDL's charitable efforts to promote reform and to preserve core constitutional principles by providing resources, training, and advocacy tools for the public, the criminal defense bar, and all those who seek to promote a fair, rational, and humane criminal justice system.

The mission of the NACDL Foundation for Criminal Justice is to preserve and promote the core values of America's justice system guaranteed by the Constitution — among them due process, freedom from unreasonable search and seizure, fair sentencing and effective assistance of counsel — by educating the public and the legal profession to the role of these rights and values in a free society.

For more information contact:

NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS®
1660 L Street NW, 12th Floor, Washington, DC 20036, Phone 202-872-8600, www.NACDL.org

NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS | NACDL FOUNDATION FOR CRIMINAL JUSTICE

This publication is available online at

www.NACDL.org/RiskAssessmentReport

II. Acknowledgements

The National Association of Criminal Defense Lawyers (NACDL) would like to thank the NACDL Foundation for Criminal Justice (NFCJ) for its support and Dr. Melissa Hamilton for authoring this report.

Special recognition goes to NACDL Past President Rick Jones, for his vision in creating and supporting NACDL's Task Force on Risk Assessment Tools and to NACDL Executive Director Norman L. Reimer for his involvement and contributions throughout the process.

This report would not have been possible without the insightful contributions and hard work of NACDL's Task Force on Risk Assessment Tools: Task Force Co-Chairs Vicki H. Young and Marvin Schechter and Task Force Members Franklin Draper, Nina J. Ginsberg, Stephen Ross Johnson, Matthew Knecht, John Philipsborn, Gail Shifman, Colette Tvedt, and CeCelia Valentine-Andrews. Thanks also goes to Daniel Weir, who served as Staff Liaison to the Task Force, and Casey Edwards for his editorial input.

We would like to thank the following NACDL staff for their careful editing and helpful suggestions: Norman L. Reimer, Executive Director; Ivan J. Dominguez, Senior Director of Public Affairs & Communications; and Cathy Zlomek, NACDL Art Director. We also thank Bonnie Hoffman, NACDL's Director of Indigent Defense Reform and Training, for her input.

During the course of its work, the task force received invaluable input from several individuals whose work has involved critical analysis of risk assessment tools and their impact: Cynthia Conti Cook, Mark Houldin, Young-Mi Lee, Scott Levy, Lisa Schreibersdorf, Alex Shalom, and Marni Jo Snyder. This acknowledgement of their generous willingness to provide input does not imply that they endorse the report or its findings and recommendations.

III. Foreword

By NACDL President Christopher W. Adams

Risk assessment tools are now used pervasively throughout the criminal legal system. They are deployed in various jurisdictions at virtually every stage, including pretrial release, charging decisions, plea bargaining, sentencing, conditions of probation and parole, suitability for early release, as well as determinations related to sexual offender registration, civil commitment, and revocation of parole or probation. Recognizing the profound implications of reliance upon these tools, in 2017 NACDL Past President Rick Jones established NACDL's Task Force on Risk Assessment Tools, chaired by Vicki H. Young and Marvin Schechter.

Initially, the Task Force sought to understand the various implications raised by the use of risk assessment instruments with the goal of formulating a policy position for consideration by NACDL's Board of Directors. As the work unfolded, and it became clear just how ubiquitous risk assessments have become, priorities shifted. While a policy might have some symbolic impact, it would have little benefit for practitioners and even less value to the clients whose lives are directly and profoundly impacted by risk assessment tools. Accordingly, the NACDL Task Force on Risk Assessment Tools commissioned Dr. Melissa Hamilton to produce a comprehensive analysis of how these tools are developed and applied. Dr. Hamilton's report represents a significant contribution to the body of scholarship and resources concerning risk assessment tools. It is an in-depth and at the same time accessible resource for practitioners, policymakers, advocates, and indeed all system actors in the nation's criminal legal apparatus. It is designed to provide the information and issue-spotting guidance necessary to properly assess various risk assessment tools, identifying their strengths and flaws.

As the report states, "This report does not take a unitary position as to whether risk assessment should be used in criminal justice. Context matters. Risk assessment may be more or less acceptable as a legal, policy, and/or ethical question depending on such matters as the nature of the decision point, the relevant rights of defendants, legal precedent, and the availability of expertise and resources. Counsel may reasonably contend for a variety of reasons that risk assessment is not suitable in a given context. Many of the points raised herein are intended to inform counsel if, despite arguments to the contrary, a risk tool is used anyway."

Dr. Hamilton's report includes "A Primer on Risk Assessment Practices" and resource-rich sections concerning "The Science Underlying Algorithmic Risk Tools," "Issues of Fairness and Bias," "Legal and Ethical Issues," and issues associated with the "Implementation of a Risk Assessment Plan."

Throughout this text, one will find "policy considerations" relevant to the risk assessment tool issue being discussed in a particular section or subsection. These considerations offer a roadmap to understanding and combatting a weakness, deficiency, defect, or worse, in a risk assessment tool, including as relates to its development and implementation. Here are just a few examples of the policy considerations that punctuate the entire report:

- “Specific discovery requests should be made by counsel to obtain training materials and codebooks used in administering the tool.”
- “Stakeholder involvement in the remediation of bias is essential for helping to ensure that the methods are appropriate for the jurisdiction. Stakeholders should be alert to biases built into risk assessment tools, both at the inception of the tool and when they are working with tools that have already been created.”
- “Mandates on data retention practices for algorithmic tool development and modification are necessary to ensure defendants have access to information about the tool, its development, its training data, the algorithm, and any updates or modifications thereto necessary to allow meaningful review.”
- “Adopting/implementing tools without trade secret protections is a crucial step toward transparency and accountability. Agencies can develop their own tools without resorting to claims of trade secrets or choose among the many available that do not claim to be proprietary.”

In the Executive Summary that follows immediately after this Foreword, the policy considerations are aggregated and annotated with page numbers and the title of the section or subsection of the report where the reader will find the discussion related to that consideration.

NACDL’s core mission includes “identifying and reforming flaws and inequities in the criminal justice system, and redressing systemic racism, and ensuring that its members and others in the criminal defense bar are fully equipped to serve all accused persons at the highest level.” This is a report that will significantly advance that mission. It is now for the reader to take in and then deploy the wealth of information, analysis, and considerations provided in this important resource.

Christopher W Adams

Christopher W. Adams
NACDL President

IV. Executive Summary

This report on risk assessment tools offers useful policy considerations throughout the body of the report. These considerations will aid those in the field representing their clients as well as policymakers and advocates in evaluating the use of various risk assessment tools, and government officials in deploying and relying upon them. For the convenience of the reader, these policy considerations are aggregated here, organized by report section, and form this Executive Summary. In addition, for each set of policy considerations, the range of pages at which the point is addressed is provided.

VII.B. A Primer on Risk Assessment Practices: Proclaimed Benefits of Risk Assessment

Policy Considerations (pages 11 to 15):

- Wherever possible, tools should incorporate protective and promotive factors that correlate with a lesser likelihood of reoffending or predict desistence.
- Risk assessments should inform placement decisions to separate low-risk from high-risk individuals.
- Low-risk outcomes may suggest a presumption of no or minimal supervisory conditions.
- Appropriate programming is best assigned according to identified criminogenic needs.

VIII.A. The Science Underlying Algorithmic Risk Tools: Tool Development

Policy Considerations (**VIII.A.3. Types of Assessments**) (pages 21 to 22):

- No presumption should exist that a proprietary or commercially developed tool performs better than government-developed or publicly available tools.
- Instruments that predict only serious offending should in most cases be adopted.
- The type of offending and type of offender the tool is designed to predict should fit the population for which an adopted tool is intended.

Policy Considerations (**VIII.A.4. Training**) (pages 22 to 23):

- The types of skills, training, and experience required to properly score the tool should be matched to the evaluators who will use it in the field.
- Adequate training on risk assessment practices in general, and on the tool adopted more specifically, is necessary. Retrainings at reasonable intervals may be appropriate for evaluators to maintain skills and when there are significant changes in the tool, its factors, or its algorithm. Material environmental changes at the site (e.g., available new programs, change in population) may dictate refresher training.
- Specific discovery requests should be made by counsel to obtain training materials and codebooks used in administering the tool.

Policy Considerations (**VIII.A.5. Collecting Information to Score**) (pages 23 to 24):

- A site should confirm that available resources will permit evaluators with regular access to all information points necessary to score the tool adopted.
- The adopting agency must verify the accuracy of information sources needed to score a tool and maintain controls to ensure improved accuracy as such sources are updated in the future.

VIII.C. The Science Underlying Algorithmic Risk Tools: Cross-Validation

Policy Considerations (pages 34 to 38):

- A tool must be cross-validated on the population and subpopulations on which it will be used, preferably before full implementation.
- Revalidation should occur at regular intervals to verify the tool's adequate performance and that the factors remain correlative with the outcome of interest.

VIII.D. The Science Underlying Algorithmic Risk Tools: Reliability

Policy Considerations (**VIII.D.1. Inter-rater Reliability**) (pages 38 to 39):

- Inter-rater reliability should be checked at regular intervals. Retraining may be necessary if reliability estimates are weak.
- If offender interviews are required to score a tool, evaluators should receive sufficient training in how to reduce interviewer bias and on culturally sensitive interview skills.

VIII.E. The Science Underlying Algorithmic Risk Tools: Communicating Risk Tool Results

Policy Considerations (**VIII.E.3. Risk Rankings Are Relative**) (pages 48 to 52):

- Agencies must make efforts to ensure that risk assessment communications are interpretable to the decision makers who receive them.
- Communication of categorical rankings should be accompanied by appropriate base rate information relevant to the population to which the defendant belongs, with 95% confidence intervals.
- Percentage estimates are preferred over relative risk as easier to understand. Still, 95% confidence intervals should also be offered.
- Communications of the likelihood of succeeding (a positive framing) is preferable for many individuals.
- The group-to-individual problem is important, and risk assessment outcomes based on group data cannot be placed onto individuals as if those outcomes were an absolute prediction.

IX.C. Issues of Fairness and Bias: How Biases May Enter Algorithms

Policy Considerations (**IX.C.6. Conflicts of Interest**) (pages 71 to 73):

- Attention to potential conflicts of interest is required in the development, modification, and validation of algorithmic tools. Cross-validation studies should be outsourced to independent researchers.
- Stakeholder involvement in the development of the algorithm is recommended so that they are aware of the potential for bias to enter algorithms and thereby provide relevant guidance. Policy choices will be required because of potential tradeoffs between bias, accuracy, individual fairness, and group fairness.

IX.D. Issues of Fairness and Bias: Efforts to Remediate Bias

Policy Considerations (pages 73 to 75):

- Stakeholder involvement in the remediation of bias is essential for helping to ensure that the methods are appropriate for the jurisdiction. Stakeholders should be alert to biases built into risk assessment tools, both at the inception of the tool and when they are working with tools that have already been created.

X.B. Legal and Ethical Issues: Due Process and Equitable Considerations

Policy Considerations (**X.B.1. The Adversarial Process)** (pages 80 to 83):

- A structure akin to mitigation services may be appropriate in terms of defense counsel offering evidence of protective and promotive factors relevant to the individual offender that may mitigate the risk score.

Policy Considerations (**X.B.5. Need for Expert Evidence**) (pages 87 to 89):

- Counsel should make use of any form of discovery that is available at the particular decision point to gain as much information as possible about the tool itself and the defendant’s scoring.
- Motions for expert witnesses at the expense of the state may be appropriate considering the nontransparency of risk algorithm processes.
- In appropriate contexts, criminal justice agencies could make available to defendants neutral subject matter experts who can explain relevant aspects of the specific algorithmic tool in terms of its development, modification, operation, validation, and biases. Funding for these experts could be included within the implementation and maintenance budget.
- Teams could develop a knowledge library to share information about specific tools, risk assessment practices, ideas about what works best, and unintended consequences of otherwise well-intentioned policies.

- More continuing legal education trainings could usefully be offered regularly and updated as risk assessment progresses into new, perhaps more technologically heavier regimes. Such training should include statistical and empirical research skills.

Policy Considerations (**X.B.6. Self-Incrimination**) (pages 89 to 90):

- Jurisdictions should ban practices in risk assessment interviews of demanding waivers of confidential information.
- Requests for confidentiality waivers should involve a right to consult with counsel.
- Questions that might elicit self-incriminating information should be excised.

Policy Considerations (**X.B.7. Validation Issues**) (pages 91 to 92):

- Authorities should carefully craft written lists of limitations that are honed to the specific tool, context (e.g., pretrial bail decision, sentencing, post-release programming), and intended population.
- Mandates on data retention practices for algorithmic tool development and modification are necessary to ensure defendants have access to information about the tool, its development, its training data, the algorithm, and any updates or modifications thereto necessary to allow meaningful review.
- Mandates on data retention practices at decision points are necessary to permit the defendant access to the data inputs, tool outcomes, and overrides that are applicable in the individual case to allow meaningful review and contest the individual score and outcome.
- Due process protections at important decision points require an evidentiary hearing and an appropriate level of discovery for the individual's assessment concerning the tool, information relied upon, and the scoring and if an override applied.

Policy Considerations (**X.B.10. Punishing the Individual for Group Behavior**) (page 94):

- Communication standards should clarify the group-based nature of the risk assessment project and that the results are relative to a group (with appropriate descriptors) and not absolute to the individual.
- Evaluators should provide 95% confidence intervals if they offer estimates of percentages normed on the developmental samples to make it clearer the variability of the statistics.

XI.A. Implementation of a Risk Assessment Program: Operational Decisions

Policy Considerations (**XI.A.1. Multidisciplinary Implementation Panel**) (pages 96 to 98):

- The jurisdiction maintaining or adopting an algorithmic tool should create, adequately fund, and sufficiently staff a multidisciplinary panel to provide oversight and a forum for debate on the many issues that can allow risk assessment practices to succeed as well as ameliorate negative consequences.

- The multidisciplinary implementation panel may include, depending on the context and decision, representatives of agency personnel, end users, defense counsel, academics, prosecution, police, forensic organizations, current or former prisoners, victims' groups, and community organizations.
- The multidisciplinary panel should engage in efforts before, during, and after implementation toward directing how the algorithmic tool is created and operates.
- The panel should consider that, at most decision points, a tool that predicts only serious offending is likely appropriate. The panel should otherwise ensure that the tool is fit for the purpose(s) of the decision it is intended to inform.
- The panel might consider if developing a tool that predicts desistence or successful reentry is desirable.
- The multidisciplinary panel should make decisions on the minimum validity levels that are acceptable and which validity and group fairness measures matter more than others.
- The multidisciplinary panel should, through open debate, make relevant decisions on how best to deal with sociodemographic characteristics and their proxies. These decisions must be weighed against cultural sensitivities and predictive abilities.
- The multidisciplinary implementation team should consider the potential for criminal history to overwhelm decisions to an unreasonable degree. Options could be to modify the algorithm to reduce reliance on criminal history measures or to build in protections within the decision framework. Limits should be placed on the use of criminal history consistent with those existing in the legal framework outside of risk assessment. Consideration should also be given to refining criminal history to include some way to factor in the age-crime curve and the progressive loss of salience of old offenses as time passes.
- A pilot study before full implementation should be conducted, if feasible.

Policy Considerations (**XI.A.2. Decision Frameworks**) (pages 99 to 100):

- The multidisciplinary panel should create a written decision framework that contains clear guidance on how the relevant decision maker/agency should use the specific tool and for what purposes.
- A risk assessment tool's outcome should never autonomously dictate a result that has negative consequence to those assessed. Instead, a tool should inform but not entirely replace a human decision maker.
- The decision framework should be clear that risk assessment results can inform but should not be used on their own to settle the ultimate issue.
- Depending on the complexity of the decision framework, training of evaluators and end users on the framework may be appropriate.

Policy Considerations (**XI.A.3. Thresholds**) (pages 100 to 102):

- The multidisciplinary team should address the placement of any thresholds, considering its goal(s), effects on predictive validity, individual fairness, and group fairness.

Policy Considerations (**XI.A.4. Communication**) (page 102):

- Risk communication practices ought to be standardized in an agency and/or jurisdiction. This might be done by the multidisciplinary implementation panel. Training on such standardization should be offered to evaluators and end users.
- Risk communication should clarify the group-based nature of assessment practices.
- Positive framing (as in the number or percentage of those who did not reoffend) may be preferable over negative framing.
- The decision framework should address how to handle missing data.

Policy Considerations (**XI.A.5. Overrides**) (page 103):

- The multidisciplinary implementation panel should consider and give clear guidance on policy and professional overrides and the discrete justifications for them.
- Discretionary overrides necessitate specific explanations in individual cases and should be subject to substantial oversight.
- Agency administrations should keep track of overrides and regularly compare override rates between evaluators in order to improve consistency in assessment.
- Risk communication to decision makers and to defendants must include transparency on whether an override was used, its form, why it was used, and overall rates of overrides. If an individual assessment is the result of an override, a statement should be required that overrides tend to reduce predictive ability of the tool.

XI.B. Implementation of a Risk Assessment Program: Accountability

Policy Considerations (**XI.B.1. Third-Party Audits**) (pages 104 to 106):

- Independent audits at regular intervals will serve interests in transparency and accountability. The body or agency adopting the risk assessment tool should ensure that appropriate funding is built in to be able to employ adequately trained and knowledgeable auditors.
- Relying on individuals, groups, or companies that are aligned with the risk assessment tool (e.g., developers, authors, consultants, employees) is not an appropriate alternative to truly independent auditors.
- An audit should include revalidating the tool on the populations for which it is scored, addressing algorithmic fairness measures, and conducting inter-rater reliability tests.
- Adopting/implementing tools without trade secret protections is a crucial step toward transparency and accountability. Agencies can develop their own tools without resorting

to claims of trade secrets or choose among the many available that do not claim to be proprietary.
- Criminal justice agencies (with multidisciplinary panel oversight) should proactively facilitate and cooperate with third-party audits by providing periodic access to data. These data sets should include individual-level data (i.e., individual offenders) with scoring information on predictive factors, outcomes (points, scores, risk bins), sociodemographic data, and recidivism data. Additional information that would be useful for auditors includes internal audit materials, training materials, codebooks, and user guides.

Policy Considerations (**XI.B.2. Adversarial Allegiance**) (pages 106 to 108):

- The decision framework should be clear when options are available (e.g., different experience tables, multiple recidivism measures), how to choose a particular option, and why to do so.
- Counsel and end users should be cognizant of the potential for adversarial bias in evaluators.

Policy Considerations (**XI.B.3. Impact Assessments**) (page 108):

- Agencies should conduct regular impact assessments after a tool's implementation.
- An impact assessment should include the following actions/elements: (a) an evaluation targeted to the goal(s) that the implementation of the tool was intended to address; (b) address the potential for external changes responding to the implementation; (c) consider intended and unintended consequences of the risk assessment regime; and (d) evaluate how risk assessment practices impact due process rights.

Policy Considerations (**XI.B.4. Data Privacy**) (page 109):

- The implementation plan should include strong protections for data privacy.

Policy Considerations (**XI.B.5. Need for Lawyering**) (pages 109 to 112):

- Legal education groups should ramp up educational offerings regarding the law, science, policy, and ethics of all things risk assessment.
- In individual cases, due process considerations mean that the defendant should have access to information on the design of the tool, validation studies, input factors, weighting, any thresholds for categorical bins, normative sample data, outputs, and override status.
- Counsel may require more time to prepare for a hearing when an algorithmic risk score was involved in the decision.

Policy Considerations (**XI.B.6. The Right to a Human Explanation**) (page 112):

Individuals have a right to a human explanation for a decision that has a substantial effect on their lives. An algorithmic risk score or categorical ranking cannot entirely replace human involvement in criminal justice decisions.

V. Introduction

Predicting recidivism using scientifically derived risk assessment tools is promoted by some as a progressive reform in criminal justice practices. The asserted value is to be able to efficiently and objectively differentiate higher- versus lower-risk individuals and for criminal justice officials to manage them accordingly. Hundreds of jurisdictions in the United States, from local to state to the federal systems, use risk tools in their decision-making. Risk tools now inform officials across decision points, such as actions related to arrest, bail, probationary conditions, sentencing, sex offender civil commitment, supervision revocation, and parole. Risk tool outcomes thereby can impose significant consequences, whether good or bad, to a criminal practitioner's clients.

Risk tool developers typically test for factors that predict the risk of offending, and they weight them according to their predictive values. Common risk factors are criminal history, age, gender, educational attainment, employment record, substance abuse, mental disorders, and relationships with criminals. In the end, developers hone their tools to a more or less complex algorithm. Here, an algorithm means a mathematical equation containing the chosen risk factors and their weights to produce a risk-based outcome. Such an outcome may be that the individual was assessed at a low, moderate, or high risk of recidivism. Alternatively, the algorithm may indicate that offenders with the same score as this defendant had an *x*% recidivism rate. Recidivism itself is variously defined, commonly using some combination of arrests, reincarcerations, supervisory revocations, technical violations, institutional disciplinary actions, and/or failures to appear for court dates.

Depending on the context, risk tools may be scored by different personnel, such as police officers in the field, probation officers when preparing pretrial recommendations or pre-sentence investigation reports, or forensic evaluators producing psychosocial evaluations. Information to score risk tools may be obtained through offender interviews (often with confidentiality waivers) or using criminal justice records, mental health reports, and other official file information.

The importance that end users, such as probation officers or judges, place on risk tool results varies. The risk outcome may be seen as one additional piece of information, may trigger some type of presumption (e.g., a presumption that low risk indicates pretrial release or high risk signifies incarceration), or may justify a more definitive consequence (e.g., high risk increases the sentencing guideline recommendation).

The guise of science has sheltered these practices from significant challenges from forensic scientists, criminal justice officials, practitioners, and other stakeholders. Consequently, risk assessment tools appear to be given too much deference when a critical eye is more befitting. This state of affairs is beginning to shift as more information is elicited, uncovered, and understood. Many complex issues exist. Risk assessment algorithms are overwhelmingly tuned to prefer higher false positive rates (erroneously classifying as high risk those who do not reoffend) than false negative rates (erroneously classifying as low risk those who do reoffend). Indeed, known error rates can exceed 50%. Tool developers have conflicts of interest, which they rarely disclose or acknowledge when reporting on the predictive performance abilities of their own tools. Despite the alleged benefit that algorithmic risk outcomes are more accurate than human judgments, biases still plague them. If the algorithm learns on already biased data, such as arrest records, those biases will become embedded into the algorithm. Also, many tools were developed using test samples consisting largely of white male adults released from maximum-security institutions. Such tools would therefore not perform as well for minorities, females, the young, or less serious offenders because of risk-relevant differences not incorporated into the risk factors chosen. The existence of human overrides likewise introduces bias. Professional overrides ordered at the discretion of individual evaluators and policy overrides instituted by officials are commonly used to overrule algorithmic outcomes. But overrides commonly increase error rates. For these various reasons, emerging evidence by independent auditors reveals disparate outcomes based on race, ethnicity, gender, and age.

Additional issues with risk assessment tools are notable. Contrary to common belief, algorithmic risk tools are not individualized predictions. They work as group-based models trained on historical data sets. The best algorithm may be able to relatively accurately identify, for example, 100 offenders as high risk, of whom 40% will recidivate. But the algorithm cannot identify which 40 individuals among the group of 100 will be the recidivists. Importantly, the vast majority of tools do not predict *serious* reoffending. Many of them count rather minor offending in their recidivism outcomes, even behaviors that do not amount to crimes. Even violent recidivism tools tend to predict the occurrence of minor assaults or threats. Further, the procedural bar to counting an act as recidivism is low. Requiring convictions is rare. Tools typically define recidivism by evidence of arrests, allegations of misconduct by supervisory officials, and/or prison disciplinary actions. Thus, these tools may be predicting a large rate of crimes that actually were not committed.

Overall, algorithmic risk assessment is a growing practice across criminal justice agencies. Yet many issues plague these tools such that care to legal and ethical issues is paramount.

VI. Overview

Risk assessment has become a mainstay in criminal justice across the United States. The following bubble commentary conceptualizes the ideals for risk assessment practices in the criminal justice system by the most ardent supporters.[1]

[1] R. Karl Hanson et al., The Council of State Governments, A Five-Level Risk and Needs System: Maximizing Assessment Results in Corrections through the Development of a Common Language 3 (2017) (internal citations omitted), https://csgjusticecenter.org/wp-content/uploads/2020/01/A-Five-Level-Risk-and-Needs-System_Report.pdf.

"Risk and needs assessments are now routinely used in correctional systems in the United States to estimate a person's likelihood of recidivism and provide direction concerning appropriate correctional interventions. Specifically, they inform sentencing, determine the need for and nature of rehabilitation programs, inform decisions concerning conditional release, and allow community supervision officers to tailor conditions to a person's specific strengths, skill deficits, and reintegration challenges. In short, risk and needs assessments provide a roadmap for effective correctional rehabilitation initiatives. When properly understood and implemented, they can help correctional organizations to provide the types and dosages of services that are empirically related to reductions in reoffending."

Council of State Governments Justice Center (2017)

Despite such a promotion, concerns have been raised whether algorithmic risk tools are fair in practice. A storm of controversy emerged when a widely publicized report claimed that a popular risk tool exhibited racial bias. In 2016, the investigative journalist group ProPublica proclaimed that the recidivism risk tool was biased against blacks.[2] ProPublica gathered data on a set of pretrial defendants who were assessed on the COMPAS tool in Broward County, Florida. ProPublica statisticians then ran the numbers. ProPublica concluded COMPAS was racist in that its algorithm produced a much higher false positive rate for blacks than whites (45% versus 24%, respectively), meaning that it overestimated high risk for blacks. COMPAS's corporate owner quickly rejected such characterization, claiming that blacks who were predicted to recidivate did so at a slightly higher rate than whites (63% versus 59%).[3]

This report engages both perspectives and will explain the seeming discrepancies in such numbers. The report begins with a primer to contextualize the development of modern risk assessment practices and explores the theoretical values and practical benefits of risk tools. A discussion addresses promising, as well as concerning, scientific issues underlying algorithmic risk assessment. It summarizes how tools are created and how their performances are evaluated.

An *algorithm* is a computation that draws in inputs to process and then produces outcomes.

This report then reviews concerns about algorithmic fairness. Controversies have emerged about how in the first place to even define algorithmic fairness. So many metrics exist that one merely needs to find the definition that fits the desired narrative that a particular tool is fair or unfair. Experience with risk tools sheds light on the various ways that bias enters what might

[2] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity* 2 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

otherwise be assumed to be value-free algorithms. A section lays out relevant legal, ethical, and policy issues. Sociodemographic factors that are scored, either directly or by proxy, raise objections. Offering algorithmic outcomes as informing critical criminal justice — based decisions face various due process and equitable challenges. Overall, despite many reasonable objections to risk assessment, jurisdictions are deploying risk-based programs across decision points. Thus, in the event algorithmic risk is in reality being used, the final section outlines best practices in implementation and maintenance of risk assessment that may attempt to ameliorate negative ramifications.

Throughout this report, the reader will encounter various policy proposals that have been put forward by stakeholders of all kinds. They are not meant to suggest any one-size-fits-all ideological or practical approach to risk-based decision-making. Context matters greatly, whereby the benefits and detriments of a risk assessment scheme may vary if applied at different decision points (from arrest, to pretrial release, to sentencing, and to post-sentence supervision). Then, risk assessment may in the individual case advantage a client given a low-risk status who is thus released. For other clients, moderate- or high-risk labels may subject them to greater penalties. In any of these cases, the need for defense counsel to understand and then to potentially promote or challenge risk assessment practices is crucial because of their significant consequences.

Risk assessment tools are unlike other types of evidence as they cannot be directly questioned or cross-examined. The unregulated nature of risk assessment, complexity of algorithms, and tendency toward opacity make it harder to access relevant information. Plus, even if algorithmic assessment performs better on average than purely human judgments, tools are plagued by significant numbers of false positives.[4] This tendency toward high false positive rates is not inevitable, though. This report details how a policy decision to simply adjust the threshold for "high-risk" outcomes can substantially reduce the false positive error rate. For these reasons, a common theme within this report's policy proposals is that if risk assessment is actually in use, there may be advantages gained from oversight by a multidisciplinary panel.

[4] Edward J. Latessa & Brian Lovins, *The Role of Offender Risk Assessment: A Policy Maker Guide*, 5 VICTIMS & OFFENDERS 203, 212 (2010).

This report does not take a unitary position as to whether risk assessment should be used in criminal justice. Context matters. Risk assessment may be more or less acceptable as a legal, policy, and/or ethical question depending on such matters as the nature of the decision point, the relevant rights of defendants, legal precedent, and the availability of expertise and resources. Counsel may reasonably contend for a variety of reasons that risk assessment is not suitable in a given context. Many of the points raised herein are intended to inform counsel if, despite arguments to the contrary, a risk tool is used anyway.

VII. A Primer on Risk Assessment Practices

Criminal justice officials have always been keen to properly manage their correctional populations to strike a reasonable balance among interests in public safety, the efficient use of limited resources, and protecting individual rights.[5] In more recent decades, correctional management has drawn on an evolving set of risk assessment practices.[6] Risk assessment here involves predicting an individual's potential for a negative criminal justice outcome, such as reoffending, supervision failure, or failure to appear. Historically, judgments on future dangerousness have been based on the gut instinct or the personal experience of the official responsible for making the relevant management decision.[7] Critics challenged those sorts of predictions as suffering from human biases.[8] The "evidence-based practices movement" is the now popular terminology to describe the turn toward using behavioral sciences research to improve offender classifications.[9] More specifically, such research informs about which factors are predictive of offending. Today, the more advanced risk assessment programs employ statistical models of prediction using automated scoring driven by algorithms.[10] The evidence-based risk movement is now in its fifth generation, as will be outlined next.

[5] Michael L. Rich, *Limits on the Perfect Preventive State*, 46 CONN. L. REV. 883, 932–33 (2014); Jay P. Singh, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 55 (2013).

[6] Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 232 (2015), http://epubs.surrey.ac.uk/id/eprint/842342.

[7] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 556 (2015).

[8] *See infra* Section VII.B.

[9] Faye S. Taxman, *The Partially Clothed Emperor: Evidence-Based Practices*, 34 J. CONTEMP. CRIM. JUST. 97, 97-98 (2018).

[10] Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 186 (2019).

A. The Evolution of Risk Tools

Evidence-based corrections practices have evolved over time such that a historical perspective unveils a generational trajectory in assessment approaches. The first generation consists of clinicians conducting unstructured or semi-structured interviews and/or file reviews to extract relevant information that, based on the professional's experience and expert knowledge, suggests recidivism risk.[11] These clinical judgments are preferred over the opinions of untrained humans. Still, clinical judgments suffer from biases such as:

> (1) ignoring or using incorrect base rates [the frequencies in which offending occurs in the populations of interest], (2) assigning suboptimal or incorrect weights to information (e.g., over-weighting "high profile" but relatively non-predictive information), (3) failing to take into account regression toward the mean [meaning that low and high scores on repeated measurements converge toward average], (4) failing to properly take into account covariation [where factors are correlated], (5) relying on illusory correlations between predictor variables and the criterion (i.e., basing decisions on the presence or absence of information that is unrelated or only weakly related to the criterion) [the criterion here meaning the recidivist act], (6) failing to acknowledge the natural bias among forensic examiners toward "conservative" judgments, defined as an increased potential for incorrect judgments of dangerousness associated with a reluctance to find someone *not* dangerous, and (7) failing to receive, and thus benefit from, feedback on judgment errors.[12]

The purely clinical judgment model produced the infamous case of Dr. Grigson, the forensic psychiatrist who testified in about 70 death penalty hearings in Texas in the 1960s and 1970s.[13] Known as "Dr. Death," he would often assert that he was 100% sure that these defendants would engage in future violence. Moving on from purely clinical assessments, the more empirically derived tools have flourished.[14]

[11] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 CRIM. JUST. & BEHAV. 1095, 1095 (2008).

[12] Eric S. Janus & Robert A. Prentky, *Forensic Use of Actuarial Risk Assessment with Sex Offenders: Accuracy, Admissibility and Accountability*, 40 AM. CRIM. L. REV. 1443, 1458 (2003) (internal citations omitted, emphasis in the original).

[13] Gregory DeClue & Denis L. Zavodny, *Forensic Use of the Static-99R*, 1 J. THREAT ASSESSMENT & MGMT. 145, 154-55 (2014).

[14] Additional information on some of the more popular algorithmic tools is available. PAMELA E. CASEY ET AL., NAT'L CTR. STATE COURTS, OFFENDER RISK AND NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS App. (2014), https://nicic.gov/offender-risk-needs-assessment-instruments-primer-courts.

Popular Risk Assessment Tools

- Static-99; Static-2002
- Violence Risk Appraisal Guide (VRAG)
- COMPAS
- Ohio Risk Assessment System (ORAS)
- Level of Service Inventory (LSI)
- HCR-20
- Public Safety Assessment (PSA)
- Federal Post Conviction Risk Assessment (PCRA)
- Offender Screening Tool (OST)

The second generation introduced actuarial instruments. These are statistical models containing factors that either are theoretically or empirically shown to correlate with recidivism. The factors are combined and scored by some derived equation.[15] They present as a "risk factorology."[16] Second generation instruments tend to be brief and efficiently scored.[17] Examples of second generation instruments are the Violence Risk Appraisal Guide (VRAG), Static-99, and the Public Safety Assessment (PSA). VRAG remains among the most popular tools to assess violent recidivism and contains twelve factors, such as age, marital status, criminal history, and psychopathy.[18] Static-99 (along with its successors) is the most widely used specifically for sex offenders to predict sexual recidivism. Static-99 contains ten static factors, five of which relate to criminal history, and several variables regard victim type, age, and cohabitation history.[19] A more recently created instrument, though it still falls within the second generation genre, is the PSA from the nonprofit Arnold Foundation. The PSA is designed for pretrial defendants and offers three scales: new criminal activity, new violent criminal activity, and failure to appear.[20] The PSA is one of the most limited in terms of containing static factors that include only age and a variety of criminal history measures.[21]

The third generation provides for structured professional judgment (SPJ) by combining the best elements of the first two generations and improving on them. SPJs provide a framework for a forensic examiner's clinical judgment and employ an actuarial model that intentionally incorporates static and dynamic factors.[22] Static risk factors normally are historical, fixed, and not

[15] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 Crim. Just. & Behav. 1095, 1095-96 (2008).

[16] Hazel Kemshall, *Crime and Risk, in* Risk in Social Science 76, 82 (Peter Taylor-Goodby & Jens O. Zinn eds., 2006).

[17] Tim Brennan et al., *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 Crim. Just. & Behav. 21, 22 (2009).

[18] Jennifer L. Skeem & John Monahan, *Current Directions in Violence Risk Assessment*, 20 Current Directions in Psychol. Sci. 38, 39 (2011).

[19] Sophie G. Reeves et al., *The Predictive Validity of the Static-99, Static-99R, and Static-2000R: Which One to Use?*, 30 Sexual Abuse 887, 887 (2018).

[20] Richard F. Lowden, *Risk Assessment Algorithms*, 19 N.C. J.L. & Tech. 221, 234-35 (2018).

[21] Arnold Foundation, *Public Safety Assessment: Risk Factors and Formula*, PSAPretrial (no date), https://www.psapretrial.org/about/factors.

[22] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 Crim. Just. & Behav. 1095, 1095-96 (2008).

amenable to interventions.[23] In contrast, dynamic factors are changeable. Dynamic factors that are criminogenic in nature present as needs the individual may have that are relevant to reducing risk. Thus, a criminogenic need is an inadequacy identified by empirical research that, if properly treated or fulfilled, diminishes the individual's likelihood of recidivism.[24] As such, criminogenic needs are suitable targets for appropriate rehabilitative programming to reduce the person's risk profile.[25] The clinical aspect of the SPJ provides structure, yet it also permits the evaluator to modify the assessment if, for instance, some idiosyncratic factor is present. This allowance overcomes the limitations of the actuarial tool whereby the human assessor may, using one's specialized knowledge, observe a risk-relevant consideration that the actuarial formula does not adequately address.[26] SPJs are not necessarily limited to licensed clinicians as evaluators. In practice, SPJs are often completed by probation officers.

The third generation thus moved from the more myopic focus on risk in the second generation tools. The correctional mindset changed as well, as the quote within the following bubble delineates.[27]

> Risk tools may inform on a variety of management issues, such as "*1) bail decisions are focused on failure to attend in court for adjudication, 2) pre-trial detention decisions require calculating public safety risk and whether a person's criminal history was sufficient to be concerned about future offending, 3) judges are concerned with whether the risk and needs should factor in sentencing decisions about the need for more restrictions or controls, and, 4) probation/parole/correctional case managers tailor the supervision plans to match the risk and needs of the offenders as an outcome. Most decisions are not about prediction of recidivism, but rather are decisions about fairness, justice, public safety, and allocation of resources to achieve public safety goals*."
>
> Taxman & Dezember (2017)

The HCR-20 is an SPJ that is among the world's most widely employed risk-needs instruments specifically for violence.[28] In summary,

[23] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 Crim. Just. & Behav. 1095, 1096 (2008).

[24] Paul Gendreau et al., *A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!*, 34 Criminology 575, 575 (1996).

[25] Paul Gendreau et al., *A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!*, 34 Criminology 575, 575–76 (1996).

[26] Stephen D. Gottfredson & Laura J. Moriarty, *Statistical Risk Assessment: Old Problems and New Applications*, 52 Crime & Delinq. 178, 181 (2006).

[27] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* Handbook on Risk and Need Assessment: Theory and Practice 22, 29 (Faye S. Taxman ed., 2017).

[28] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 Behav. Sci. & L. 554, 555 (2018).

> [the] HCR-20 is so-named for its inclusion of 20 risk factors in Historical, Clinical, and Risk management domains. The instrument contains 10 historical, largely static, risk factors that fall into three general categories (problems in adjustment or living, problems with mental health, and past antisocial behavior) and 10 potentially changeable, dynamic risk factors. Five of these concern current clinical status such as negative attitudes and active symptoms of major mental illness (the Clinical scale), and five concern future situational risk factors such as lack of plan feasibility and treatment noncompliance (the Risk Management scale).[29]

The Level of Service Inventory-Revised (LSI-R) is an SPJ instrument that is one of the most commonly used generic risk-needs tools across American criminal justice agencies.[30]

> [The LSI-R] contains 54 items rationally grouped according to the following 10 subcomponents representing different risk/need areas: Criminal History, Education/Employment, Finances, Family/Marital, Accommodations, Leisure/Recreation, Companions, Alcohol/Drug, Emotional/Personal, and Attitude/Orientation. Items are scored as either present or absent, based on a semistructured interview and review of available file information, and subsequently summed to yield a total score. Higher scores reflect a greater risk of recidivism and need for intervention.[31]

In the next iteration, fourth generation assessments supplement the risk-needs combination with responsivity principles and provide a longer-term perspective on case management spanning from intake through case closure.[32] The responsivity aspect concerns "tailoring case plans to the individual characteristics, circumstances, and learning style of each offender."[33]

The federal probation system developed its Post Conviction Risk Assessment (PCRA) as a fourth generation, software-based tool that identifies risk factors and barriers to treatment.[34] The PCRA scores a variety of static and dynamic factors, including education, employment, substance abuse, family problems, and individuals' attitudes toward supervision and behavioral change.[35] The PCRA includes a self-survey component for defendants to complete that is then scored to elicit their criminal thinking styles.

[29] Laura S. Guy et al., *Assessing Risk of Violence Using Structured Professional Judgment Guidelines*, 12 J. FORENSIC PSYCHOL. PRAC. 270, 272 (2012).

[30] Memorandum from Vera Inst. of Justice to Del. Justice Reinvestment Task Force 4 (Oct. 12, 2011).

[31] David J. Simourd & P. Bruce Malcolm, *Reliability and Validity of the Level of Service Inventory-Revised Among Federally Incarcerated Sex Offenders*, 13 J. INTERPERSONAL VIOLENCE 261, 264 (1998).

[32] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 CRIM. JUST. & BEHAV. 1095, 1096 (2008).

[33] WINNIE ORE & CHRIS BAIRD, NAT'L COUNCIL ON CRIME & DELINQUENCY, BEYOND RISK AND NEEDS ASSESSMENTS 8 (2014), http://nccdglobal.org/sites/default/files/publication_pdf/beyond-risk-needs-assessments.pdf.

[34] Christopher T. Lowenkamp et al., *The Federal Post Conviction Risk Assessment (PCRA): A Construction and Validation Study*, 10 PSYCHOL. SERVICES 87, 88 (2013).

[35] James L. Johnson et al., *The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)*, 75(2) FED. PROB. 16, 26 app. 2 (2011).

The Correctional Offender Management Profiles for Alternative Sanctions (COMPAS), one of the best-known fourth generation tools,[36] is a "web-based tool designed to assess offenders' criminogenic needs and risk of recidivism. Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision, and case management of offenders."[37] Reflecting the progress made in the fourth generation, COMPAS distinguishes itself. "COMPAS has several modules: risk/needs assessment, criminal justice agency decision tracking, treatment and intervention tracking, outcome monitoring, agency integrity, and programming implementation monitoring. The risk assessment component addresses four basic dimensions: violence, recidivism, failure to appear, and community failure."[38]

Within the fourth generation, some tools incorporate factors that reduce the potential of future dangerousness.[39] These may entail protective factors that moderate (lessen) the salience of a risk factor, or promotive factors that predict desistence (i.e., not engaging in or ceasing the possible behavior flagged by the risk factor).[40] Risk assessment had moved beyond simply risk prediction by more forthrightly addressing an orientation toward risk management.[41] Unfortunately, the evidence to date does not clearly show that third or fourth generation instruments perform better at predicting recidivism than second generation tools.[42] However, they have better utility with the needs component.

Advances in behavioral sciences, the availability of big data, and improvements in statistical modeling have ushered in a wave of algorithmic risk assessment tools. The second through fourth generation tools use more or less sophisticated algorithms in their actuarial models. Indeed, algorithms have been creeping further into the criminal justice system as a general rule, and there may be no limit to them. To the extent that human behavior is perceived as predictable and calculable by a set of identified factors, algorithms can theoretically take over more and more criminal justice decisions. Algorithms are thus displacing human engagement.[43]

[36] Tracy L. Fass et al., *The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools*, 35 CRIM. JUST. & BEHAV. 1095, 1097 (2008).

[37] NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS 1 (2013).

[38] Sheldon X. Zhang et al., *An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures*, 60 CRIME & DELINQ. 167, 172 (2014).

[39] PAMELA E. CASEY ET AL., NAT'L CTR. STATE COURTS, OFFENDER RISK AND NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS 11 (2014), https://nicic.gov/offender-risk-needs-assessment-instruments-primer-courts.

[40] John Monahan & Jennifer Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CRIM. PSYCHOL. 489 (2016).

[41] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 PSYCHOL. ASSESSMENT 1456 (2019).

[42] Faye S. Taxman, *Risk Assessment: Where Do We Go From Here*?, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 271, 274 (Jay P. Singh et al. eds., 2018).

[43] Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC DAVIS L. REV. ONLINE 261, 267 (2019), https://hyp.is/go?url=https://lawreview.law.ucdavis.edu/online/vol52/52-online-Hamilton.pdf&q=user:qdr@hypothes.is.

Human–Algorithm Interaction: Options

- A decision support system in which the algorithm aids the human decision maker
- Human in-the-loop where humans have some involvement in the resulting decision
- Completely delegating decisions to algorithmic tools

Computer modeling may also be replacing humans in the development of risk tools themselves. A fifth generation of risk assessment scales is just being recognized that employs machine learning in the creation of the algorithm and produces real-time risk estimates.[44] Pennsylvania is at the forefront of piloting machine learning forecasts for parole decisions.[45] An initial study found no impact of this machine learning tool on the rate of parole releases but did detect changes in the mix of offenders paroled and consequent reductions in recidivism from parolees.[46] In any event, these new algorithmic risk protocols have been variously depicted as "smart" and "data-driven," offering "big data analytics" and a type of "algorithmic governance."[47]

A potential new wave of science-informed risk comes from neuroscience and is reliant on neuroimaging to predict violent reoffending specifically. Unlike the five generations just mentioned, the hope is that neuroscience can detect *causal* mechanisms between brain function/structure and aggression.[48] Neuroimaging studies are currently underway using inmates in American institutions to develop this science with a promise for more individualized predictions and targeted medical interventions.[49] Early indications reveal certain types of brain disorders or altered connectivity have some relationship to violent acts.[50] Neuroprediction is beyond the scope of this report, but active work in this area is worthy of mention here.

B. Proclaimed Benefits of Risk Assessment

Risk factorology recognizes that the crime for which one is suspected of committing may

[44] Brandon L. Garrett & John Monahan, *Judging Risk*, 108 Calif. L. Rev. 439, 451 (2020).

[45] Richard Berk, *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*, 13 J. Experimental Criminology 193, 193 (2017).

[46] Richard Berk, *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*, 13 J. Experimental Criminology 193, 193 (2017).

[47] Kelly Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates*, 23 Theoretical Criminology 453, 454 (2019).

[48] Russell A. Poldrack et al., *Predicting Violent Behavior: What can Neuroscience Add?*, 22 Trends Cognitive Sci. 111 (2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5794654.

[49] John Philipsborn, *A Basic Assessment Toolbox: Aiming for Adequate Lawyering During the Spread of Risk Assessments*, Champion 18, 21-22 (Jan./Feb. 2020).

[50] Carl Delfin et al., *Prediction of Recidivism in a Long-Term Follow-up of Forensic Psychiatric Patients: Incremental Effects of Neuroimaging Data*, 14(5) PLOS One (2019), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217127.

"Sentencing and parole authorities around the world are turning to behavioral science for guidance in triaging offenders into those who require rehabilitative interventions in institutional settings and those who can more effectively be treated in the community."

Monahan (2018)

provide limited information about the danger the person poses to the community.[51] In general, risk assessment purportedly offers the ability to reduce mass incarceration by diverting low-risk defendants from prison while targeting greater supervision and services to those at higher risk.[52] "Data-driven algorithmic decision making may enhance overall government efficiency and public service delivery, by optimizing bureaucratic processes, providing real-time feedback and predicting outcomes."[53] With such a tool in hand, criminal justice officials can more consistently input relevant data and receive software-produced risk classifications.[54] The more advanced models that incorporate needs, protective factors, and promotive factors may better serve the interests and futures of defendants in that they recognize and hopefully support the potential for humans to change their behaviors and to improve their lives.[55]

Algorithmic risk assessments now drive how officials triage offenders into placements, supervision levels, and programming.[56] Algorithmic risk tools are often lauded for reducing the arbitrariness of,[57] and errors produced by, subjective human judgments.[58] Human decision-making is replete with explicit and implicit biases.[59] The formalized structure of the new formats and their actuarial computations help mitigate the influence of such cognitive biases on decisions.[60] Human judgments may vary for reasons that computer programs do not, such as by

[51] Nathan James, Cong. Res. Serv., *Risk and Needs Assessment in the Criminal Justice System* (Oct. 13, 2015), https://digital.library.unt.edu/ark:/67531/metadc795663/m1/1/high_res_d/R44087_2015Oct13.pdf.

[52] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 PSYCHOL. SCI. 206, 206 (2016).

[53] Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Challenges*, 31 PHIL. & TECH. 611, 612 (2018).

[54] J. Stephen Wormith, *Automated Offender Risk Assessment*, 16 CRIMINOLOGY & PUB. POL'Y 281, 285 (2017).

[55] Ralph C. Serin & Caleb D. Lloyd, *Integration of the Risk Need, Responsivity (RNR) Model and Crime Desistance Perspective: Implications for Community Correctional Practice*, 7 ADVANCING CORRECTIONS 37, 38 (2019).

[56] John Monahan, *Preface: Recidivism Risk Assessment in the 21st Century, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS xxiii, xxiii (Jay P. Singh et al. eds., 2018).

[57] Monika Zalnieriute et al., *From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making, in* CAMBRIDGE HANDBOOK ON THE LAW OF ALGORITHMS (Woodrow Barfield ed., forthcoming 2020), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3380072_code722134.pdf?abstractid=3380072&mirid=1.

[58] Jeffrey C. Singer et al., *A Convergent Approach to Sex Offender Risk Assessment, in* THE WILEY-BLACKWELL HANDBOOK OF LEGAL AND ETHICAL ASPECTS OF SEX OFFENDER TREATMENT AND MANAGEMENT 341, 343 (Karen Harrison & Bernadette Rainey eds., 2013).

[59] Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 186 (2019) (citing studies).

[60] SARAH PICARD-FRITSCHE ET AL., CENTER FOR COURT INNOVATION, DEMYSTIFYING RISK ASSESSMENT: KEY PRINCIPLES AND CONTROVERSIES 2 (2017).

simply being tired, hungry, or distracted.[61] Humans can disregard rules through creative machinations while algorithms (unless driven by unsupervised machine learning techniques) cannot reject any strict rules embedded therein.[62]

Automated tools are perceived as reducing the potential for human corruption.[63] Thus, "[v]endors promote these models to the public and to the agencies that use them as the answer to human bias, arguing that computers cannot harbor personal animus or individual prejudice based on race, gender, or any other legally protected characteristic."[64] Algorithmic risk prediction is also thought to exceed a human brain's ability to efficiently and rationally process multiple data points.[65]

A side benefit is that the record-keeping required by the tools in terms of data collection, inputs, and scoring helps bring greater accountability to decisions informed by them.[66] Automation limits personal discretion and, representing evidence-based practices, could improve the defensibility of decisions as well.[67]

Using algorithmic learning to create these tools in the first place may also be preferable. Algorithms are just better at identifying personal and social correlations within individuals and across people that pertain to human activity.[68] Algorithmic processing can find patterns in behavior that humans may not be capable of cognitively replicating. Further, algorithms may detect predictive factors that appear counterintuitive.[69] It could be, for example, that committing crimes of greater severity such as homicide may not correlate with a higher risk of serious reoffending.

Algorithmic outputs can also help structure any human aspect of decision-making. An Indiana Supreme Court decision praised risk assessment tools as aiding judges in a sentencing context to "more effectively evaluate and weigh several express statutory sentencing considerations such as criminal history, the likelihood of affirmative response to probation or short term imprisonment, and the character and attitudes indicating that a defendant is 'unlikely to commit another crime.'"[70]

[61] R. Barry Ruback et al., *Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis*, 80 FED. PROB. 47, 47 (2016).

[62] Monika Zalnieriute et al., *From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making, in* CAMBRIDGE HANDBOOK ON THE LAW OF ALGORITHMS (Woodrow Barfield ed., forthcoming 2020).

[63] Monika Zalnieriute et al., *From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making, in* CAMBRIDGE HANDBOOK ON THE LAW OF ALGORITHMS (Woodrow Barfield ed., forthcoming 2020).

[64] Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 186 (2019).

[65] Angéle Christin et al., *Courts and Predictive Algorithms* 1 (2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf.

[66] R. Barry Ruback et al., *Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis*, 80 FED. PROB. 47, 47 (2016).

[67] HAZEL KEMSHALL, RISK IN PROBATION PRACTICE (2019).

[68] Betsy Anne Williams et al., *How Algorithms Discriminate Based on Data They Lack*, 8 J. INFO. POL'Y 78, 82-83 (2018).

[69] Saul Levmore & Frank Fagan, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. (forthcoming 2020).

[70] Malenchik v. State, 928 N.E.2d 564, 574 (Ind. 2010).

While identifying high-risk persons to incapacitate may assuage the public's fear, there are special benefits to screening out low-risk persons. Studies indicate that low-risk pretrial defendants have better chances of success if released pending trial.[71] Pretrial defendants who remain incarcerated are more likely to be victimized in jail, plead guilty, and receive longer sentences than those who are released.[72] Pretrial incarceration is also linked to job loss and family disruption, which are known predictors of reentry failure.[73] Still, releasing pretrial defendants with burdensome conditions does not benefit low-risk offenders. Indeed, pretrial release conditions increase the risk of failure for low-risk individuals, with the potential exception of applicable mental health treatment.[74]

Similar patterns are observed for post-conviction defendants after release who are low risk. Providing low-risk offenders with overly restrictive conditions of supervision, requiring intrusive programming, or placing them with higher-risk inmates can often be counterproductive in that those interventions actually increase recidivism rates in low-risk populations.[75] Release with many conditions provides more opportunities for supervision failure, and overprogramming may interfere with self-corrective efforts.[76] Intensive supervision is also poorly tolerated by low-risk offenders because it often entails mixing with high-risk offenders who share their criminal attitudes/behaviors, while also disrupting low-risk offenders' connections with prosocial contacts.[77] A relevant study of federal offenders on supervised release compared the outcomes of low-risk offenders before and after the initiation of a policy reducing the number of supervisory contacts and found a lower arrest rate after the policy implementation.[78] Informed by such research, assessments can serve the system by influencing decisions that have the effect of "[r]educing social, economic, and family costs associated with inappropriate, and often counter-productive, interventions with low-risk offenders."[79]

As a result of the foregoing benefits, risk assessment practices in criminal justice have attracted prominent advocates, such as the American Bar Association, the Conference of State Court Administrators, the Conference of Chief Justices, and the National Association of

[71] Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in the Federal Court*, 73(2) FED. PROB. 3, 6 (2009), https://www.uscourts.gov/file/22893/download.

[72] Melissa Hamilton, *The Biased Algorithm*, 56 AM. CRIM. L. REV. 1553, 1557 (2019), http://epubs.surrey.ac.uk/id/eprint/852008.

[73] Laura I. Appleman, *Justice in the Shadowlands: Pretrial Detention, Punishment, & the Sixth Amendment*, 69 WASH. & LEE L. REV. 1297, 1320 (2012).

[74] Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in the Federal Court*, 73(2) FED. PROB. 3, 6 (2009), https://www.uscourts.gov/file/22893/download.

[75] Christopher T. Lowenkamp et al., *The Risk Principle in Action: What Have We Learned from 13,676 Offenders and 97 Correctional Programs*, 51 CRIME & DELINQ. 1, 13 (2006).

[76] Nathan James, Cong. Res. Serv., *Risk and Needs Assessment in the Federal Prison System* 15 (July 10, 2018), https://justiceroundtable.org/wp-content/uploads/2018/07/Risk-and-Needs-Assessment-in-the-Federal.pdf.

[77] Thomas H. Cohen, *The Supervision of Low-Risk Federal Offenders: How the Low-Risk Policy has Changed Federal Supervision Practices without Compromising Community Safety*, 80(1) FED. PROB. 3, 3 (2016).

[78] Thomas H. Cohen, *The Supervision of Low-Risk Federal Offenders: How the Low-Risk Policy has Changed Federal Supervision Practices without Compromising Community Safety*, 80(1) FED. PROB. 3, 8 tbl. 5 (2016).

[79] PAMELA E. CASEY ET AL., NAT'L CTR. STATE COURTS, OFFENDER RISK AND NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS 6 (2014).

Counties.[80] More recently, the adoption of risk assessment practices to drive criminal justice reform is one of the few major issues that has received bipartisan political support.[81] Nonetheless, there remain staunch critics of risk assessment for multiple and valid reasons, which will be set forward throughout this report.

Policy Considerations:

Wherever possible, tools should incorporate protective and promotive factors that correlate with a lesser likelihood of reoffending or predict desistence.

Risk assessments should inform placement decisions to separate low-risk from high-risk individuals.

Low-risk outcomes may suggest a presumption of no or minimal supervisory conditions.

Appropriate programming is best assigned according to identified criminogenic needs.

C. Decision Points

The National Institute of Corrections, affiliated with the Department of Justice, advocates risk-needs assessment at every stage in the criminal justice process.[82] Jurisdictions may mandate that risk assessment tools be used through policy or by law,[83] though these directives exhibit disparate levels of sophistication and clarity.[84] Ohio has adopted a risk assessment scheme for use across criminal justice decision points.[85] The Model Penal Code has embraced risk assessment at sentencing, at least with sufficient validation and reliability.[86] Virginia, Utah, and Pennsylvania have been at the forefront of engaging algorithmic tools for sentencing purposes.[87] Reports exist of some type of informal use of risk assessment by judges in sentencing in Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Washington, and Wisconsin.[88] Overall, risk assessment data are utilized in sentencing in about 20 states and for parole decisions in more than half of states.[89]

[80] Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CAL. L. REV. 439 (2020).

[81] Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L. J. 2218, 2222 (2019).

[82] Julia Angwin et al., *The Legal System Uses an Algorithm to Predict if People Might be Future Criminals: It's Biased Against Blacks*, MOTHER JONES (May 23, 2016 5:16 PM).

[83] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 555 (2018).

[84] John Philipsborn, *A Basic Assessment Toolbox: Aiming for Adequate Lawyering During the Spread of Risk Assessments*, CHAMPION 18, 19 (Jan./Feb. 2020).

[85] Nathan James, Cong. Res. Serv., *Risk and Needs Assessment in the Criminal Justice System* 4-5 (Oct. 13, 2015), https://digital.library.unt.edu/ark:/67531/metadc795663/m1/1/high_res_d/R44087_2015Oct13.pdf.

[86] Model Penal Code: Sentencing §6.06(5) (Final Draft 2017).

[87] John Monahan & Jennifer Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CRIM. PSYCHOL. 489, 495 (2016).

[88] Julia Angwin et al., *The Legal System Uses an Algorithm to Predict if People Might be Future Criminals: It's Biased Against Blacks*, MOTHER JONES (May 23, 2016, 5:16 PM).

[89] Jodi L. Viljoen et al., *Do Risk Assessment Tools Help Manage and Reduce Risk of Violence and Reoffending? A Systematic Review*, 42 LAW & HUM. BEHAV. 181, 181 (2018).

Algorithmic risk tools have been incorporated into pretrial systems across states such as Arizona, Kentucky, and New Jersey.[90] The Annie E. Casey Foundation developed the Risk Assessment Instrument for juvenile justice, and it has been deployed in more than 300 jurisdictions across 39 states.[91]

The foregoing consider just a few of the areas of risk assessment adoption. Algorithmic risk assessment informs a host of other criminal justice decision points. Figure 1 provides a list.

Figure 1: Decision Points in Risk Assessment Practices

Decision Points for Risk Assessment

- Arrest
- Diversion
- Bail
- Deferred adjudication
- Plea negotiations
- Probation conditions
- Community supervision level
- Supervisory contact frequency
- Probation revocation
- Juvenile transfer to adult systems
- Programming
- Competency
- Insanity defense
- Parole
- Supervised release conditions
- Reentry services
- Parole revocation
- Drug court decisions
- Reentry court decisions
- Sex offender registration
- Sex offender civil commitment
- Death penalty
- Inmate security classification
- Institutional placement
- Solitary confinement
- Earned good time

Clearly, risk assessment has permeated across the criminal justice system. A select few of the current tools are specifically designed for discrete decision points. For example, the Public Safety Assessment (introduced earlier) is focused on pretrial release decisions, and Virginia and Pennsylvania developed instruments to inform specifically on sentencing decisions. Nonetheless, most tools are used across multiple decision points, whether or not they were designed or intended by the developers for such purposes. The lack of regulatory or other legal controls in this area means that, in reality, this type of "off-label" use is rampant.

[90] Glen J. Dalakian II, *Open the Jail Cell Doors, Hal: A Guarded Embrace of Pretrial Risk Assessment Instruments*, 87 FORDHAM L. REV. 325, 342 (2018).

[91] Angèle Christin et al., *Courts and Predictive Algorithms* 3 (Oct. 27, 2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf.

Risk assessment tools may or may not be designed for discrete decision points. So-called "off-label" use of tools for multiple purposes not intended by their developers is in practice quite common. No clear legal framework or industry norm currently exists to regulate off-label use.

The next section describes the common development processes for modern algorithmic tools.

VIII. The Science Underlying Algorithmic Risk Tools

The previous discussion outlining the generational development referred to statistical models beginning in the second generation as actuarial in nature. As the sophistication level has increased and with the dominance of computer aided modeling (e.g., machine learning, artificial intelligence), more recent projects aimed at improving preexisting tools or developing new ones are likely to be referred to as algorithmic in nature. The term algorithm better captures these technological advancements. The following is a summary of how the more complex risk tools are created.

A. Tool Development

With growing acceptance and use, the number and scope of risk assessment tools have only expanded. There are likely hundreds of tools in use across the globe. Tool developers have emerged along many fronts.

1. Risk Tool Developers

Risk assessment is a competitive industry, as the following quote suggests.[92]

> *"Recidivism prediction is ubiquitous. Everybody's doing it. There is an enormous academic and professional literature. Unprecedented private sector involvement has occurred in designing and marketing instruments and providing services to government."*
>
> Tonry (2014)

Tools are created by governmental agencies, nonprofit institutes, and for-profit businesses.[93] Developers may also be individuals or groups of individuals working in the field with forensic

[92] Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 FED. SENT'G REP. 167, 167 (2014).

[93] Leon Neyfakh, *You Will Commit a Crime in the Future: Inside the New Science of Predicting Violence*, BOSTON GLOBE, Feb. 20, 2011,

science expertise. Many instruments are proprietary and require payment for their use, while others are in the public domain.[94] On occasion, the private/public divide shifts. Previously, the Arnold Foundation's Public Safety Assessment was considered confidential and required monies to access. The Arnold Foundation now makes its tool for pretrial purposes freely available to jurisdictions and has published its algorithms.[95]

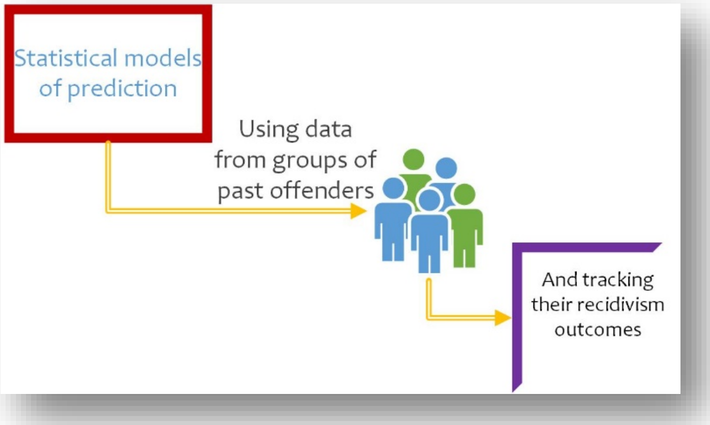
Despite some tools requiring sometimes substantial investments by users to access them, there is no evidence that proprietary models consistently perform better at predictions than those that are freely available or developed by governmental agencies.[96]

2. *Stages of Development*

Some of the early tools were informed by a literature review of criminological theories of what may motivate criminal offending. Tools emerging recently are more likely to be developed through the statistical analyses of data sets of offenders.

Algorithmic risk tools basically rely on aggregate statistics derived from historical samples of offenders (often referred to as training, developmental, or normed samples). These underlying data sets vary in size from dozens to over a million cases. Developers study the statistical relationships between the information points available in the data set and the risk outcome of interest (e.g., a recidivist event). Figure 2 presents a simple visual of an early stage of the development process.

Figure 2: Early Tool Development


Statistical models of prediction
Using data from groups of past offenders
And tracking their recidivism outcomes

Generally, for the basic risk assessment tools, developers are interested in finding statistically significant predictors of recidivism. Researchers then select from the strong predictors and assign appropriate weights considering some factors will have greater predictive value than others.[97]

http://www.boston.com/bostonglobe/ideas/articles/2011/02/20/you_will_commit_a_crime_in_the_future.

[94] Susan Turner & Julie Gerlinger, *Risk Assessment and Realignment*, 53 SANTA CLARA L. REV. 1039, 1045 (2013).

[95] Arnold Foundation, *About*, PSAPRETRIAL, https://www.psapretrial.org/about/factors.

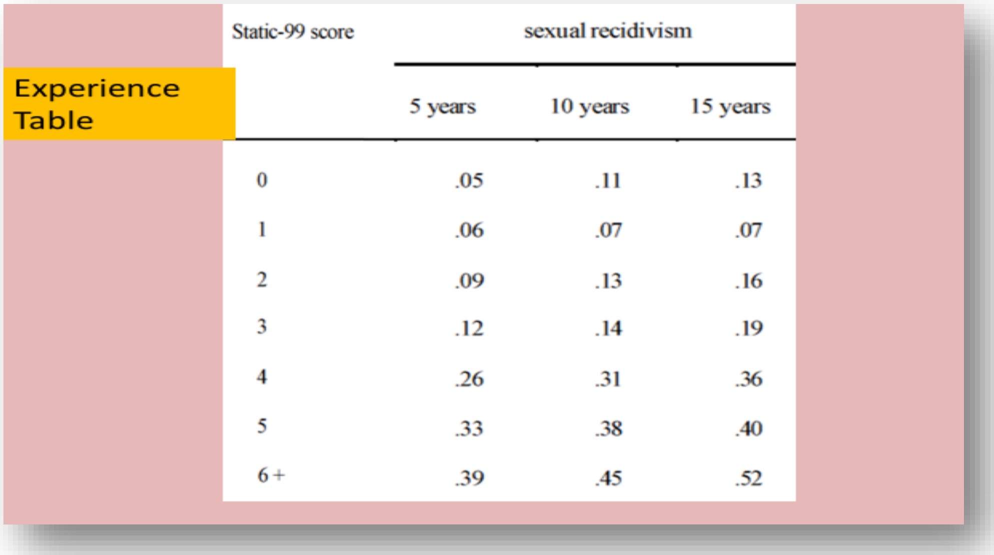
[96] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 3, 15-17 tbls. 1.4-1.6 (Jay P. Singh et al. eds., 2018).

[97] Joanna Amirault & Patrick Lussier, *Population Heterogeneity, State Dependence and Sexual Offender Recidivism: The Aging Process and the Lost Predictive Impact of Prior Criminal Charges over Time*, 39 J. CRIM. JUST. 344, 344 (2011).

Thus, the algorithm is created that drives the tool's scoring system. A table of estimated probabilities of the outcome occurring is often created to match to final scores. This is called an experience table since it is based on the observed rates of recidivism from the testing samples.

Figure 3 presents an experience table from Static-99 to use for illustration purposes.

Figure 3: Static-99 Experience Table


Experience Table

Static-99 score	sexual recidivism		
	5 years	10 years	15 years
0	.05	.11	.13
1	.06	.07	.07
2	.09	.13	.16
3	.12	.14	.19
4	.26	.31	.36
5	.33	.38	.40
6+	.39	.45	.52

Static-99 utilizes three different follow-up periods: 5, 10, and 15 years. These matter as the longer an individual's time at risk, the more opportunity there is to reoffend. Thus, at any given score, observed recidivism rates will increase over time. In Figure 3, the left-hand column represents Static-99 scores. The original Static-99 scale ranged from 0 to 10 points. The instrument lumps together scores of 6 and above into one bin because the historical data did not justify distinguishing scores that met or exceeded the 6 points. The next three columns signify the recidivism rates in decimals. Thus, Figure 3 would convey that, of the subjects who were assigned a score of 5 in the testing sample, 33% were observed to have sexually recidivated at five years, 38% were observed to have sexually recidivated at 10 years, and 40% sexually recidivated in a 15-year period.

In sum, developers of basic risk instruments typically (a) locate a historical data set, (b) test factors that correlate with recidivism, (c) combine highly relevant factors, (d) provide applicable weights to create the scoring algorithm, (e) consider combining scores together, and (f) provide an experience table of estimated probabilities of the outcome.

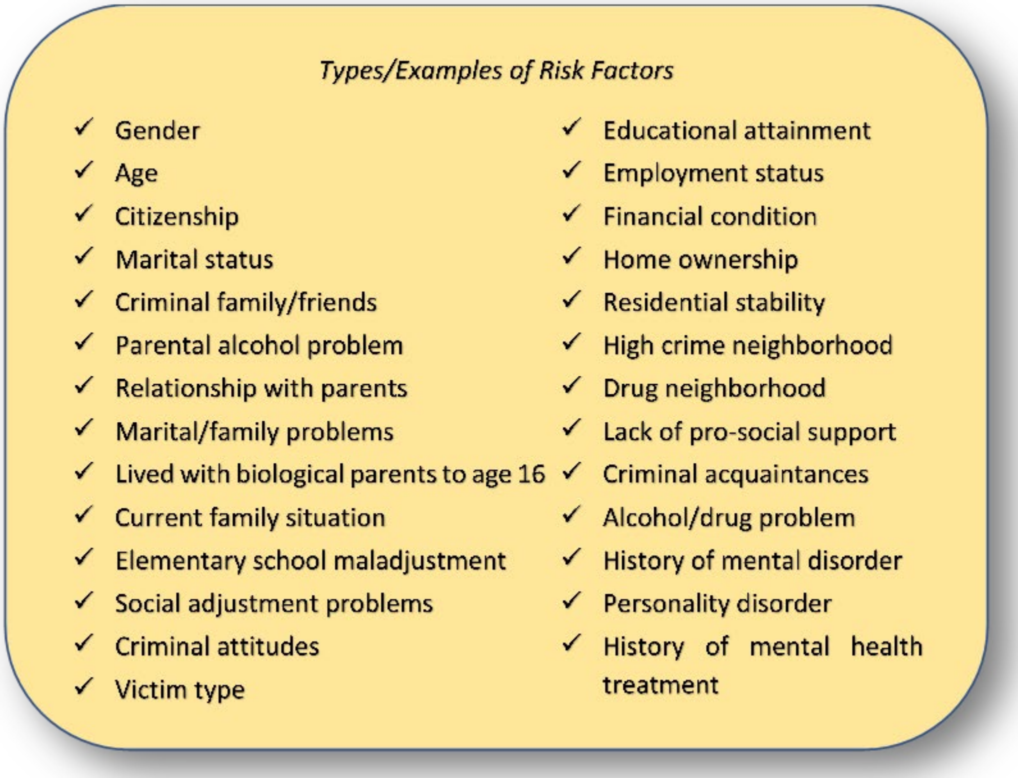
Developers of actuarial risk tools at times pool together risk groupings, referred to as risk bins, based on point totals.[98] The Static-99 strategy groups scores of 0 to 1 as Low risk, 2 to 3 as Low-Moderate, 4 to 5 as Moderate-High, and 6-plus as High risk.[99]

Figure 4 conveys examples of risk and needs factors actually present in criminal justice tools today.

[98] Jay P. Singh, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 57 (2013).

[99] Katherine E. McCallum et al., *The Influence of Risk Assessment Instrument Scores on Evaluators' Risk Opinions and Sexual Offender Containment Recommendations*, 44 CRIM. JUST. & BEHAV. 1213, 1220 (2017).

Figure 4: Risk and Needs Factors



Types/Examples of Risk Factors

- ✓ Gender
- ✓ Age
- ✓ Citizenship
- ✓ Marital status
- ✓ Criminal family/friends
- ✓ Parental alcohol problem
- ✓ Relationship with parents
- ✓ Marital/family problems
- ✓ Lived with biological parents to age 16
- ✓ Current family situation
- ✓ Elementary school maladjustment
- ✓ Social adjustment problems
- ✓ Criminal attitudes
- ✓ Victim type
- ✓ Educational attainment
- ✓ Employment status
- ✓ Financial condition
- ✓ Home ownership
- ✓ Residential stability
- ✓ High crime neighborhood
- ✓ Drug neighborhood
- ✓ Lack of pro-social support
- ✓ Criminal acquaintances
- ✓ Alcohol/drug problem
- ✓ History of mental disorder
- ✓ Personality disorder
- ✓ History of mental health treatment

3. *Types of Assessments*

Tools are designed to predict some form of criminal justice failure. A subset of tools is designed to predict discrete types of failures (e.g., violent behavior, failure to appear, institutional infraction). Certain tools are tailored to specific types of offenders, such as those known to have committed a violent offense, a sexual offense, or domestic abuse. Notably, no algorithmic tool is yet available for terrorists because of their extreme rarity, unique motivations, and the lack of large groups of released terrorists to study.[100] Tools may likewise be limited by selected sociodemographic characteristics, such as being designed exclusively for those with a confirmed criminal history, males, juveniles, or those with specified mental disorders.[101]

A little-known fact is that most tools are not limited to predicting serious crime.

[100] *See generally* Melissa Hamilton, *A Threat Assessment Framework for Lone-Actor Terrorists*, 70 FLA. L. REV. 1319 (2018).

[101] Leon Neyfakh, *You Will Commit a Crime in the Future: Inside the New Science of Predicting Violence*, BOSTON GLOBE, Feb. 20, 2011.

Risk tools typically do not limit themselves to predicting *serious* offending.

Tools often include minor infractions in their recidivism definitions, including those that do not rise to the level of constituting crime. The following are some examples:

- traffic stops or municipal ordinance violations (California Pretrial Assessment Tool)[102]
- negative reports from a parole supervisor or a parole violation (including drinking or taking non-prescribed drugs) (Self-Appraisal Questionnaire)[103]
- "antisocial behavior, such as institutional misconduct or breach of supervision" (the LSI family)[104]
- technical failures, pretrial failures, disciplinary problems while incarcerated (COMPAS)[105]
- serious or nonserious infractions while incarcerated (STRONG-R)[106]
- evidence of any violent act defined to include "(1) a person engaged in an act or omission (2) with some degree of willfulness that (3) caused or had the potential to cause (4) physical or serious psychological harm to (5) another person or persons" (HCR-20^{v3})[107]

As the last bullet suggests, even violent risk tools tend to count simple assaults. Note also that HCR-20 includes threats of serious psychological harm as violence.

i *Policy Considerations:*

No presumption should exist that a proprietary or commercially developed tool performs better than government-developed or publicly available tools.

Instruments that predict only serious offending should in most cases be adopted.

The type of offending and type of offender the tool is designed to predict should fit the population for which an adopted tool is intended.

4. Training

Risk tool scoring has in practice been completed by a variety of personnel, such as police officers, probation officers, social workers, psychologists, mental health personnel, parole

[102] Chelsea Barabas et al., *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns* 1 n. 2 (July 17, 2019), https://dam-prod.media.mit.edu/x/2019/07/16/TechnicalFlawsOfPretrial_ML%20site.pdf.

[103] Wagdy Loza, *Self-Appraisal Questionnaire (SAQ): A Tool for Assessing Violent and Non-Violent Recidivism*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 165, 172 (Jay P. Singh et al. eds., 2018).

[104] J. Stephen Wormith & James Bonta, *The Level of Service (LS) Instruments*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 117, 117 (Jay P. Singh et al. eds., 2018).

[105] Tim Brennan & William Dieterich, *Correctional Offender Management Profiles for Alternative Sanctions (COMPAS)*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 49 (Jay P. Singh et al. eds., 2018).

[106] Zachary Hamilton, *The Static Risk Offender Needs Guide-Revised (STRONG-R)*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 199, 203 (Jay P. Singh et al. eds., 2018).

[107] Kevin S. Douglas et al., *Historical-Clinical-Risk Management-20, Version 3 (HCR-20^{v3}): Development and Overview*, 13 INT'L J. FORENSIC MENTAL HEALTH 93, 100 (2014).

officials, or forensic nurses.[108] Still, care must be taken that evaluators have the appropriate level of skill, training, access to necessary data, time, and resources available to properly score the factors incorporated therein. Tools that incorporate mental health factors, for instance, likely should require sufficient education and experience in psychological diagnoses. (The exception here may be if the tool expects mental health diagnoses to be scored from available medical file information.)

Some developers require that users have a certain professional background, specified amount of instruction, and/or certification for use. Other developers make suggestions on the evaluators' experience and training. Obviously, a relevant attribute here is that owners of proprietary tools have some ability to dictate and manage minimal qualification standards. Developers of tools placed in the public domain, though, maintain little control over who is using their tool and how much training and experience they have with it.

Rigorous training on how to score the tool is also important to realize the advantages of regularity in assessments. Tools often come with codebooks to standardize the scoring of individual items, though the lengths and the extent of details therein vary. Tools can contain factors that are not so objective on their face but for which the developers have crafted precise meanings as dictated in codebooks or user guides. As an example, the VRAG's factor of "elementary school maladjustment" does not in itself carry clear normative meaning.

ℹ *Policy Considerations:*

The types of skills, training, and experience required to properly score the tool should be matched to the evaluators who will use it in the field.

Adequate training on risk assessment practices in general, and on the tool adopted more specifically, is necessary. Retrainings at reasonable intervals may be appropriate for evaluators to maintain skills and when there are significant changes in the tool, its factors, or its algorithm. Material environmental changes at the site (e.g., available new programs, change in population) may dictate refresher training.

Specific discovery requests should be made by counsel to obtain training materials and codebooks used in administering the tool.

5. Collecting Information to Score

Evaluators in the field must gather information in response to the factors within the tool adopted. Developers tend to dictate the means to collect the information necessary to score their tools. Data collection may be based on one or more of the following: criminal justice records, mental health records, institutional files, or other administrative records; offender self-assessment surveys; offender interviews; and information provided by professionals with knowledge of the offender.

[108] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 427 (2016).

Effective prediction requires quality data be input into the algorithm. Information of dubious quality just results in "garbage in, garbage out."

Policy Considerations:

A site should confirm that available resources will permit evaluators with regular access to all information points necessary to score the tool adopted.

The adopting agency must verify the accuracy of information sources needed to score a tool and maintain controls to ensure improved accuracy as such sources are updated in the future.

B. Accuracy and Validation

A reason for the interest in creating evidence-based models, and continuing to improve on them, is to increase the accuracy of algorithmic risk predictions over human judgments. To determine how accurate any tool may be in correctly predicting a criminal justice failure, the tool's abilities must be tested. The general idea is that best practices dictate that a tool be validated on the population(s) on which it is intended to be used.

1. Validity Measures

Validity simply means the extent to which a test properly reflects the concept it is designed to reflect.[109] Here, validity asks whether the tool adequately measures the risk of the type(s) of criminal justice failure that the tool is designed to predict. A tool that is said to be "validated" is not necessarily one that is highly accurate. Validation is minimally achieved if the tool predicts recidivism at a rate statistically greater than chance. This equates to the proverbial state of being better than a flip of the coin. As previously discussed, the criminal justice outcome of interest is often a recidivist act, supervision failure, or failure to appear. For ease of reference, the discussion that follows will simply refer to the relevant outcome as a recidivist act.

A tool may be said to be "validated" simply because it predicts recidivism at a rate statistically greater than chance.

Developers will commonly divide their original data set into two. One set is the training sample. The other is reserved to validate the algorithm that is created from the training sample.

The narrative below delves into specific measures that address the validity of an instrument. The purpose of providing a detailed explanation herein is threefold. First, these measures are important because they address various ways to determine how accurate a tool is and to elicit what types of errors it produces. Second, these measures will matter again when this report addresses algorithmic fairness later on. The third reason is to provide a basis for understanding the confusion underlying the ProPublica debate with the owner of COMPAS and how these

[109] MICHAEL G. MAXFIELD & EARL BABBIE, RESEARCH METHODS IN CRIMINAL JUSTICE AND CRIMINOLOGY 127 (2d ed. 1998).

organizations could study the same data and come to opposite conclusions whether the same tool was biased against blacks.

Multiple measures related to validity are available to judge the capabilities of an assessment tool. These measures evaluate distinguishable aspects of the tool. Importantly, there exist two main aspects to validity: (1) its discriminative ability and (2) its calibration. *Discrimination* reflects how well the tool distinguishes recidivists from non-recidivists.[110] Discrimination represents the tool's relative accuracy in terms of the ability to differentiate recidivists from non-recidivists. Discrimination is retrospective in nature as it is calculated after the recidivists and non-recidivists have been identified.[111]

In contrast, *calibration* concerns how accurate the tool statistically estimates recidivism, and it measures the tool's absolute predictive accuracy.[112] Calibration is forward-looking; it measures how well the tool predicts future recidivism. Hence, discrimination and calibration each offers distinct contributions to judging a tool's validity. As a result, a tool may vary in how well it meets either of these metrics.

> A scale that ranks well, but systematically overestimates or underestimates risk might have good discriminative properties but be poorly calibrated to the population under examination; in contrast, a very simple scale (e.g., one that merely divided offenders into ever violent/never violent, or male/female groups) might be very well-calibrated but have only modest discriminative validity.[113]

Thus, validity can be divided into two aspects: its ability to discriminate and its calibration. In turn, discrimination and calibration can themselves be subdivided in terms of there being precise calculations available to address them.

a. The Contingency Table

Popular measures of discrimination and calibration rely on numbers contained in what is referred to as a 2×2 contingency table. Figure 5 shows this contingency table. The "2×2" depiction references the fact that the table contains two rows and two columns.

[110] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[111] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[112] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[113] Philip D. Howard, *The Effect of Sample Heterogeneity and Risk Categorization on Area Under the Curve Predictive Validity Metrics*, 44 CRIM. JUST. & BEHAV. 103, 105 (2017).

Figure 5: The 2 × 2 Contingency Table

		Outcome		
		Recidivist	Non-Recidivist	
Assessment	High Risk	True Positives (TP)	False Positives (FP)	*Positive Predictive Value (PPV)*
	Low Risk	False Negatives (FN)	True Negatives (TN)	*Negative Predictive Value (NPV)*
		True Positive Rate (TPR)	*True Negative Rate (TNR)*	

The contingency table is filled with numbers derived from studying a population of offenders who were scored on a risk tool and their recidivism outcomes known. The table also requires that one constrict a risk tool's potential assessments into two categories: a likely recidivist or a likely non-recidivist. Often these categories use the term *high risk* for the likely recidivist and *low risk* for the likely non-recidivist. These terms (*high risk* and *low risk*) are not necessarily synonymous with a tool's actual use of these terms. Instead, these are simply employed here to divide the population assessed into two groupings. For instance, if a particular tool offered a range of scores from 1 to 10, the researcher must choose a specific score along that continuum as the cutoff. The researcher may choose 6 such that those who score between 1 and 5 are combined into the "low-risk" group, while those scoring between 6 and 10 are collected into the "high-risk" grouping for purposes of completing the 2×2 contingency boxes.

A contingency table provides a host of information. True Positives (TP) are the number who are judged as high risk and did commit a recidivist act. False Positives (FP) are those at high risk who did *not* commit a recidivist act. In turn, True Negatives (TN) are the number who were correctly classified as low risk as they did not commit a recidivist act. The final number within the table concerns the False Negatives (FN), representing the low-risk individuals who committed a recidivist act.

One method for judging the overall accuracy of a tool is to combine its correct assessments overall by adding TN and TP and dividing that sum by the entire population (N). Table 1 includes the overall accuracy statistic just mentioned but also contains various calculations for several discrimination and calibration items.

Table 1: Discrimination and Calibration Measures

Measure	*Calculation*
Overall Accuracy	$\frac{TP+TN}{N}$
True Positive Rate (TPR)	$\frac{TP}{TP+FN}$
True Negative Rate (TNR)	$\frac{TN}{TN+FP}$
False Positive Rate (FPR)	$1-TNR$
False Negative Rate (FNR)	$1-TPR$
Positive Predictive Value (PPV)	$\frac{TP}{TP+FP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$
False Discovery Rate (FDR)	$1-PPV$
False Omission Rate (FOR)	$1-NPV$

Let us first look at the true positive rate (TPR) and the true negative rate (TNR), which represent high-risk and a low-risk discrimination metrics, respectively.[114] The TPR is alternatively titled "sensitivity" in the field of statistics and represents the accuracy rate for the recidivists.[115] The TNR is alternatively titled "specificity" and represents the accuracy rate for the non-recidivists.[116]

> [These measures derive from the field of] military signal detection, where a more sensitive signal is less specific. Sensitivity is the true positive rate: the proportion of actual events that are identified as events. Specificity is the true negative rate: the proportion of actual nonevents that are identified as nonevents. Signal detection involves a trade-off: If the detector is tuned to be more sensitive, it will detect a greater proportion of correct targets (e.g., spot enemy ships), but it will also be less specific and therefore produce more false alarms (e.g., mistake dolphins for enemy ships). Translated to the prediction of recidivism, the "detector" is a risk assessment scale, which can be reported either as a score on the risk scale or a probability of recidivism associated with that score.

[114] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 9 (2013).

[115] Kristian Linnet et al., *Quantifying the Accuracy of a Diagnostic Test or Marker*, 58 CLINICAL CHEMISTRY 1292, 1296 (2012).

[116] Kristian Linnet et al., *Quantifying the Accuracy of a Diagnostic Test or Marker*, 58 CLINICAL CHEMISTRY 1292, 1296 (2012).

> It can be tuned by varying the binary threshold above which those assessed are considered to be at heightened risk of recidivism.[117]

The TPR and TNR provide discrimination statistics and thus are *retrospective* in nature. They take known recidivists and non-recidivists to determine whether they had been predicted to reoffend. Then both can be flipped to ascertain error rates in discrimination. The False Negative Rate is the reciprocal of TPR, while the False Positive Rate is the reciprocal of TNR.

In comparison, two metrics that more appropriately measure *prospective* predictive accuracy—and thereby are more important to practitioners who are interested in the predictive validity of risk tools—are the positive predictive value (PPV) and negative predictive value (NPV).[118] The PPV represents the probability that a high-risk score was correct (i.e., it is the proportion of high-risk predictions who became recidivists).[119] The NPV then is the proportion of those predicted at low risk who did not recidivate. The PPV is a high-risk calibration measure, while the NPV is a low-risk calibration measure.[120] The False Discovery Rate is the reciprocal of the PPV, while the False Omission Rate is the reciprocal of NPV.

In the 2×2 table in Figure 5, then, the columns are discrimination measures where the recidivist act is already known to have occurred (or not). The rows are calibration measures where the recidivist act is not yet known when the risk prediction is made. Understanding how these measures are distinguishable is important given that a tool can perform well on one or more of them while showing poor results on others.

Perhaps an analogy to testing in the medical care profession may help clarify these distinctions. A diagnostic test in medicine attempts to determine if a person now has a particular disease. Discrimination is analogous to diagnosing that a person is a recidivist in that his (additional) crime exists. A prognostic test in medicine predicts the likelihood a person will in the future get a specific disease. Calibration, likewise, predicts whether a person will in the future be a recidivist in that a crime will likely occur.

The distinction between discrimination and calibration explains the controversy about the tool COMPAS. In 2016, the investigative journalist group ProPublica kickstarted a public debate on the topic when it proclaimed that the tool COMPAS was biased against blacks.[121] ProPublica obtained the data through Freedom of Information Act requests and public websites. Recall that ProPublica concluded COMPAS was racist in that its algorithm produced a much higher false positive rate for blacks than whites (45% versus 24%, respectively), meaning that it overestimated high risk for blacks. COMPAS's corporate owner, Northpointe, quickly rejected such characterization. After running its own statistical analyses on the same data set ProPublica had compiled, Northpointe statisticians asserted that their results demonstrated COMPAS

[117] Philip D. Howard, *The Effect of Sample Heterogeneity and Risk Categorization on Area Under the Curve Predictive Validity Metrics*, 44 CRIM. JUST. & BEHAV. 103, 105 (2017).

[118] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 12 (2013).

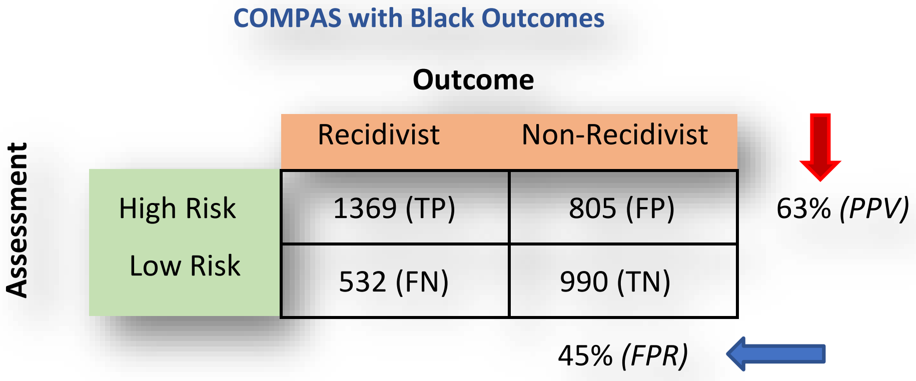
[119] ROBERT H. RIFFENBURGH, STATISTICS IN MEDICINE 206 tbl. 10.3 (3d ed. 2012).

[120] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 11 fig. 1 (2013).

[121] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016).

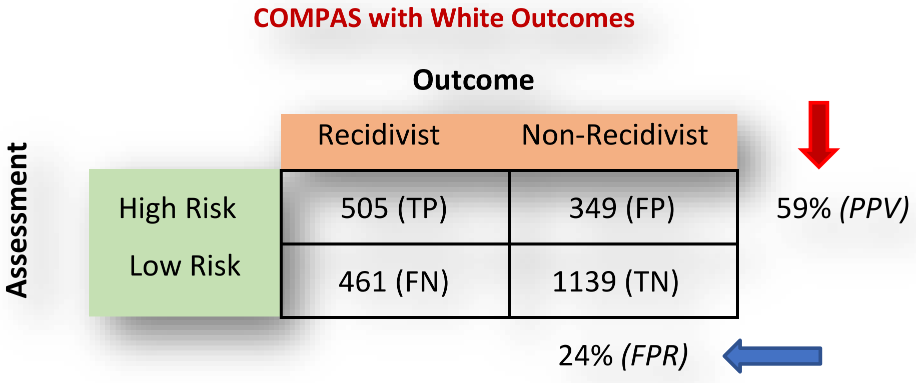
outcomes achieved predictive parity for blacks and whites.[122] Northpointe reported that black defendants who were predicted to recidivate did reoffend at a "slightly" higher rate than whites (63% versus 59%, respectively). ProPublica and Northpointe analyzed the same data set yet were simply addressing different aspects of the same 2×2 contingency table. ProPublica calculated the tool's discrimination ability (the columns) while Northpointe focused on calibration (the rows). Table 2 shows that the result was that both groups' calculations were correct at the same time, yet because they relied on different measurements, their inconsistent conclusions were given statistical support.

Table 2: Validity of COMPAS With Race



COMPAS with Black Outcomes

Assessment	Outcome		
	Recidivist	Non-Recidivist	
High Risk	1369 (TP)	805 (FP)	63% *(PPV)*
Low Risk	532 (FN)	990 (TN)	
		45% *(FPR)*	



COMPAS with White Outcomes

Assessment	Outcome		
	Recidivist	Non-Recidivist	
High Risk	505 (TP)	349 (FP)	59% *(PPV)*
Low Risk	461 (FN)	1139 (TN)	
		24% *(FPR)*	

From Table 2 one can see that ProPublica drew on the FPR statistics from the right-hand columns to declare that the false positive rate for blacks was 45% compared with 24% for whites. Northpointe, though, selected the positive predictive value from the top row to declare that COMPAS predicted recidivism for blacks at 63% and at 59% for whites.

b. Area Under the Curve

Another common metric in the risk assessment literature for testing the discriminatory ability of a tool is called the area under the curve (AUC). The AUC is derived from a statistical plotting of

[122] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity* 2 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

true positives and false positives across a risk tool's rating system.[123] More specifically, the AUC is a discrimination index that represents the probability that a randomly selected recidivist received a higher risk classification than a randomly selected non-recidivist.[124] The size of the risk scale differential between them is irrelevant; as long as the risk classification of the recidivist is even minimally higher, it will count positively toward the AUC.[125] AUCs range from 0 to 1.0, with .5 indicating no better accuracy than chance, and a 1.0, meaning perfect discrimination (i.e., all recidivists were classified higher than all non-recidivists).[126]

Risk assessment scholars often refer to AUCs of .56, .64, and .71 as the thresholds for small, medium, and large effect sizes, respectively.[127] An effect size simply refers to the magnitude of the relationship between two variables. Using this guidance, an AUC of .56 means that recidivists are rated higher than non-recidivists in 56% of cases. This .56 is better than chance, but only slightly so, which is why it is labeled a small effect size. The large effect size indicates significant strength; but, still, it means that the tool correctly discriminated 71% of the time, leaving a 29% discrimination error rate. Notably, agreement on the strength of AUCs is not universal.[128] A more conservative conceptualization is that AUCs between .60 and .69 are poor, .70 to .79 are fair, .80 to .89 are good, and over .90 are excellent.[129]

The AUCs of many known validation studies across tools hover around .70, meaning that the tools correctly distinguish recidivists from non-recidivists about 70% of the time. The reason for such consistency is that tools tend to measure the same types of factors: criminal history, criminal lifestyle, antisocial personality, and alcohol/mental health issues.[130] There is also the potential that there is some natural limit to predicting human behavior. With this in mind, an observer may wonder why experts staunchly warn about the need to conduct validation studies on new populations and about the potential for biases against particular groups. An issue called publication bias is known to occur but not how often. Publication bias refers to the tendency for researchers not to publicize results of studies that do not have significant results.[131] This is also called the file drawer problem in that a study with poor results will simply be stuck in a file and forgotten. Here, it is quite possible that validation studies with poor AUCs simply have been

[123] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 15 (2013).

[124] Jay P. Singh et al., *Measurement of Predictive Validity in Violence Risk Assessment Studies*, 31 BEHAV. SCI. & L. 55, 64 (2013).

[125] Philip D. Howard, *The Effect of Sample Heterogeneity and Risk Categorization on Area Under the Curve Predictive Validity Metrics*, 44 CRIM. JUST. & BEHAV. 103, 107-08 (2017).

[126] Martin Rettenberger et al., *Prospective Actuarial Risk Assessment: A Comparison of Five Risk Assessment Instruments in Different Sexual Offender Subtypes,* 54 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 169, 176 (2010).

[127] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 12 (2017).

[128] Jay P. Singh, *Five Opportunities for Innovation in Violence Risk Assessment Research*, 1 J. THREAT ASSESSMENT & MGMT. 179, 181 (2014).

[129] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[130] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 PSYCHOL. ASSESSMENT 1456 (2019).

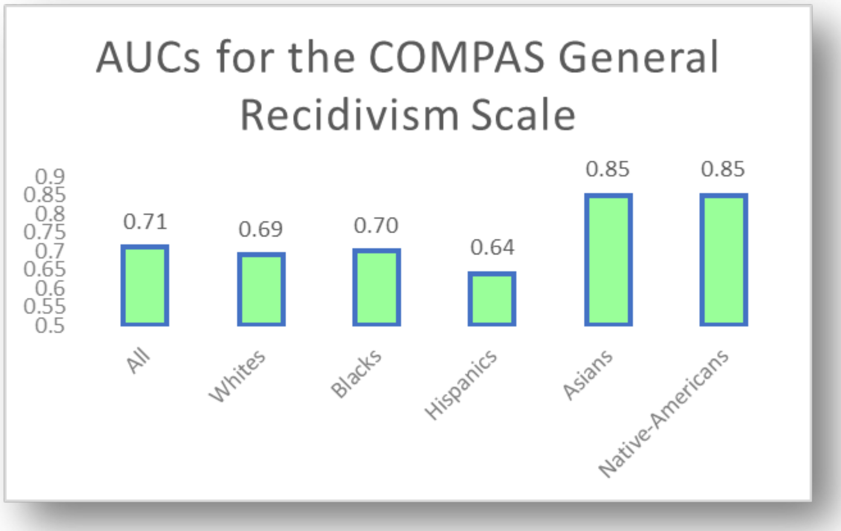
[131] Kerry Dwan et al., *Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review*, 8(7) PLOS ONE (2013), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0066844.

hidden.

Further, most (known) validation studies provide AUC metrics on entire populations rather than by subgroups within the population. This may obscure differences in AUC levels for subgroups, particularly if they represent small percentages of the entire sample.

The Reporter has previously conducted various statistical analyses on a live data set using the COMPAS tool and offenders assessed on it in Broward County, Florida, in 2013 to 2014. This is the same data set collected by ProPublica that was mentioned at the beginning of this report. ProPublica was transparent enough to post its data set on the web for other researchers to explore. Results from those analyses will be offered at various places herein for illustration purposes. Figure 6 provides an example of AUCs for different racial/ethnic groups for the COMPAS tool using the Broward County data set.

Figure 6: AUCs by Group for COMPAS



As shown in Figure 6, the overall AUC is .71 (a randomly chosen recidivist would have scored higher than a randomly chosen non-recidivist 71% of the time). However, the AUCs varied by race/ethnicity, from a low of 64% discriminatory ability for Hispanics up to 85% for Asians and Native Americans. This illustrates how certain reporting customs can obscure group differentials. Moreover, these results suggest group bias, a concept that will be addressed below.

Despite its frequent reference in the risk assessment literature, the AUC has serious limitations and thus cannot present a holistic portrait of a tool's abilities.[132] Unfortunately, the AUC is too commonly misinterpreted as measuring calibration accuracy; but a higher AUC does not mean more accurate prospective prediction.[133] As well, the AUC cannot calculate how well

[132] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 16-18 (2013).

[133] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 16 (2013).

an instrument selects those at moderate or high risk.[134] The AUC could achieve a large effect size even if no recidivists were ranked as high risk. To use a hypothetical, suggest a tool ranks individuals along a scale from 1 to 10, with selected groups of 1 to 4 labeled low risk, 5 to 7 as medium risk, and 8 to 10 indicating high risk. The AUC for such a tool would actually reflect perfect accuracy (AUC = 1.0) where all recidivists were classified as level 2 and all non-recidivists as level 1. Yet, in a scale from 1 to 10, there is very little distinction between the scores, and all had been labeled low risk.

Additionally, the AUC does not distinguish between types of errors. Whether they are predominantly false positives or false negatives is simply not picked up in this single statistic. But this likely matters to officials who generally have an interest in whether they prefer a higher rate of false positives versus false negatives.[135] Another flaw is that AUC accuracy rates between groups may be comparable, but the type of error may differ between groups, such as one group having a higher rate of false positives, yet another a higher rate of false negatives.[136]

In sum, an assertion about achieving a certain AUC level is not a ringing endorsement to support a claim that a tool is well validated. For readers who still have difficulty grasping the statistical concepts discussed so far, perhaps a more accessible approach is desired. More enterprising writers have conceptualized many of these terms using baseball analogies.[137]

Risk assessment accuracy statistics can be explained through analogies to baseball terms.

2. *Accuracy (and Error) Rates Are Changeable*

Importantly, the foregoing measures (e.g., TPR, TNR, PPV, NPV) are malleable. The reason is that they require a chosen cut-point to distinguish low risk versus high risk. Changing the cut-point alters these statistics. "Choosing a very high threshold for classification would imply higher number of forecasted low risks, and therefore more false negatives. On the other hand, choosing a very low threshold would result in more false positives and few false negatives."[138]

Let us employ the tool COMPAS in a hypothetical. COMPAS ranks risk on a scale of 1 to 10. Suggest the selected cut-point is a score of 5 whereby those scoring 1 to 4 are grouped into a low-risk designation while scores of 5 to 10 present as high risk for purposes of these calculations. Doing so will permit a statistician with sufficient outcome data to calculate the cells in the 2×2 contingency table. If one were, instead, to choose the cut-point as 8 (scores 1 to 7 as low risk and

[134] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 Behav. Sci. & L. 8, 17 (2013).

[135] Jorge M. Lobo et al., *AUC: A Misleading Measure of the Performance of Predictive Distribution Models*, 17 Global Ecology & Biogeography 145, 146 (2008).

[136] Solon Barocas et al., *Big Data, Data Science, and Civil Rights* (2017), https://arxiv.org/pdf/1706.03102.

[137] *See generally* Christopher P. Marett & Douglas Mossman*, From Ballpark to Courtroom: How Baseball Explains Risk Assessment*, 47 Psychiatric Annals 443 (2017).

[138] Garima Siwach & Shawn D. Bushway, *Adaption of Risk Tools to the Employment Context*, *in* Handbook on Risk and Need Assessment: Theory and Practice 475, 494 (Faye S. Taxman ed., 2017).

8 to 10 as high risk) and rerun the calculations, the set of TPR, TNR, PPV, and NPV statistics will differ, in some cases significantly, solely because of the shift in cut-point. More specifically, by increasing the threshold for high risk, the TPR and NPV will decrease while TNR and PPV will increase.[139]

We will show this in Table 3 using the entire Broward County data set with COMPAS and adding in the discrimination error rates.

Table 3: Changing Cut-Points for COMPAS

Measure	Cut-Point 5		Cut-Point 8
True Positive Rate	62%		30%
True Negative Rate	70%		91%
False Positive Rate	30%		9%
False Negative Rate	38%		70%
Positive Predictive Value	63%		74%
Negative Predictive Value	69%		61%

By increasing the threshold, the accuracy of the (retrospective) true positive rate falls by more than half (from 62% to 30%), while the accuracy of the (retrospective) true negative rate increases (from 70% to 91%). As expected, a higher bar for high risk increases the false negative rate (the inverse of TPR) from 38% to 70% while reducing the false positive rate (the inverse of TNR) from 30% to 9%. Raising the cut-point increases accuracy in the (prognostic positive) predictive value (from 63% to 74%), while decreasing for the negative predictive value (from 69% to 61%).

In other words, by moving the cut-point higher, the diagnostic and prognostic abilities change, but in opposite directions. The classification of known recidivists is substantially worse while the classification of the known non-recidivists improves. In contrast, the predictive ability improves for those at high risk while worsening for those predicted low risk. As tools are meant to be predictive, arguably, the latter measures (PPV and NPV) matter more.

Notice the trade-offs that can be made if one is more concerned with false positives than false negatives (or vice versa). A simple shift in the cut-point threshold can significantly influence the predictive accuracy.

No industry standards exist for which cut-points to use. For purposes of these statistics, it is often the developers who may have suggestions on cut-points for their own tools and/or the independent scientists conducting validation studies who choose them. Yet cut-points fail to represent empirically driven numbers. Instead, such choices should result from stakeholder

[139] Seena Fazel, *The Scientific Validity of Current Approaches to Violent and Criminal Risk Assessment, in* Predictive Sentencing: Normative and Empirical Perspectives 197, 200 (Jan W. de Keijser et al. eds. 2019).

decisions on what constitutes high versus low risk and what likelihood (percentage) is acceptable or not for any/all of the relevant accuracy measures.[140] This topic is discussed further below.

C. Cross-Validations

Tool developers generally will check that their tool performs better than chance with the training sample. Regardless of how well the tool functions on the test data, it is not advisable to simply transport that tool across to new populations and settings. Developers intentionally construct their algorithms to be the best fit for the training sample. As a result, the accuracy statistics are likely to be at their highest with the training data. Additionally, any test performed in controlled lab conditions, such as in the evolution of a tool outside of any real-world application, may not do as well in the field, as indicated in the following quote.[141]

> *"[I]nstruments are likely to produce better results under closely supervised research conditions than in regular clinical practice, and that in studies in which an instrument is first developed, its predictive accuracy is optimized and is inevitably less when repeated on an independent sample."*
>
> Szmukler & Rose (2013)

Significant issues exist for other reasons with any presumption that a recidivism assessment tool is generalizable outside of the training samples. "[T]here is no way to tell in the development sample how much of the observed relation between the variables and recidivism is due to underlying associations that will be shared in new samples and how much is due to unique characteristics of the development sample."[142]

Criminal acts and their correlates can vary depending on personal characteristics and experiences, times, geographies, sites, environments, and circumstances.[143] The following quote provides a good summary of suggested types of cross-validating factors:

> [The] predictive efficacies of all tools must be eventually subjected to repeated empirical validation with client groups that differ in demographic characteristics (e.g., age, gender, socioeconomic status, ethnicity), level and type of past violence (e.g., criminal histories, sexual vs. nonsexual offenders), psychiatric diagnosis (e.g., presence of personality disorder, psychosis), intervention received (e.g., treated vs. untreated), the specific criterion being predicted (e.g., violent vs. nonviolent behavior

[140] Discussion at Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[141] George Szmukler & Nikolas Rose, *Risk Assessment in Mental Health Care: Values and Costs*, 33 BEHAV. SCI. & L. 125, 130 (2013).

[142] Gina M. Vincent et al., *The Use of Actuarial Risk Assessment Instruments in Sex Offenders, in* SEX OFFENDERS: IDENTIFICATION, RISK ASSESSMENT, TREATMENT, AND LEGAL ISSUES 70, 81 (Fabian Saleh et al. eds., 2009).

[143] Keith Soothill, *Sex Offender Recidivism*, 39 CRIME & JUST. 145, 176 (2010).

> or different types of violent behavior), environmental setting (e.g., clients residing in institutions vs. the community), countries of origin of the research, and so forth.[144]

Population drivers of crime may distinguish geographic areas with different rates of crime from one another (e.g., poverty rates, unemployment, social upheaval), but they are generally not represented in tool predictors. Another consideration is that validation metrics and the success of a tool's deployment may depend on the jurisdiction's courtroom culture, policing habits, prosecutorial practices, and community interests.[145] Moreover, recidivism risk tools have generally incorporated variables found to be *associated with* reoffending; researchers did not intend to prove *causation*. The final variables are not, then, shown to cause recidivism.

Officials occasionally disregard the advisory against transporting tools to new locations without pretesting to ensure proper validation. Recent experience informs that "the application of risk knowledge is often haphazard: jurisdictions frequently deploy pre-existing screening tools in settings for which they were neither designed nor calibrated, and legal and correctional officials frequently do not understand the actuarial technologies on which they base their decisions."[146]

Validation is context-dependent. Hence, best practices dictate that a tool be validated on the specific population for which it will be used in a real-life setting.[147] Cross-validation testing can confirm (or refute) whether the algorithm performs adequately there. Also, questions are being raised in the scientific and policy communities that an otherwise "reasonable algorithm" may fail to result in fair and equitable treatment of diverse populations.[148] A cross-validation can test how well the algorithm performs across groups, such as minorities and women.

One court stands out, albeit not within the United States, in acknowledging the importance of cross-validation and that the failure to do so may present a legal impediment. The Canadian Supreme Court in 2018 precluded officials from using a particular tool (COMPAS) on an Indigenous Canadian prisoner because there was no evidence that the tool had been validated on that subpopulation.[149] The justices observed that "substantive equality requires more than simply equal treatment" as treating groups identically may itself produce inequalities.[150] The judges there may be correct to question whether tools are appropriate for native populations. Other tools have failed to validate on indigenous offenders in the few studies that address those groups.[151]

Revalidation of a tool may even be necessary on the same population that previously had

[144] Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740, 741 (2010).

[145] Chelsea Barabas et al., *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns* 4 (July 17, 2019).

[146] Seth J. Prins & Adam Reich, *Can we Avoid Reductionism in Risk Reduction?*, 22 THEORETICAL CRIMINOLOGY 258, 260 (2018) (internal citations omitted).

[147] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing Standard 3.10 (1999).

[148] OSONDE OSABA & WILLIAM WELSER IV, RAND CORP., AN INTELLIGENCE IN OUR IMAGE: THE RISKS OF BIAS AND ERRORS IN ARTIFICIAL INTELLIGENCE 19 (2017), https://www.rand.org/pubs/research_reports/RR1744.html.

[149] Ewert v. Canada, 2018 S.C.R. 30, para. 66 (S.C.C. June 13, 2018).

[150] *Id.* at para. 54.

[151] R. Karl Hanson et al., *Assessing the Risk and Needs of Supervised Sex Offenders*, 42 CRIM. JUST. & BEHAV. 1205, 1219 (2015).

been shown to have a sufficient level of accuracy. Even if core recidivism risk factors remain viable, over time some risk factors may change in their predictive strength. For example, at one point not having a landline in the home was a predictive factor in a pretrial context for failure to appear.[152] Clearly, with telecommunication changes, the salience of a landline has eroded such that it is no longer a justifiable risk factor. Revalidations can assist by identifying factors that are no longer sufficiently predictive.

Experience tables of observed recidivism rates at different risk bins or scores may also need to be updated if the base rates of offending in that population have shifted. The significant drop in rates of violence and sex offending in the United States from the 1990s onward is a reason that many studies using new samples tend to show a mismatch with the violence and sexual risk assessment tools developed on dated samples.[153] A well-known case highlighting the issue of changing base rates is with the popular risk assessment tool Static-99.[154] The original norms were based on sex offenders released in the 1960s to 1980s. When the developers collected new data on released sex offenders, they found that recidivism rates dropped considerably. Static-99 developers offered some suggestions for this result.

> Possible explanations that have been proposed include demographic factors (e.g., aging population, increased obesity, reliance on medications such as Prozac and other serotonin-affecting agents), cultural factors (e.g., changing mores regarding sexuality, increased awareness about sexual assault leading to greater vigilance and supervision of children), and criminal justice factors (e.g., offender treatment, increased supervision, deterrent/incapacitation effects of longer sentences).[155]

Consequently, Static-99 developers created and issued new experience tables with updated recidivism rates. (Still, the original table remains publicly accessible, and evidence exists that at least some evaluators continue to use the original table anyway.)[156]

Another word of caution about such experience tables is prudent here. A tool's experience table created from its testing samples may simply not be replicated in other samples. And the rates of reoffending within the same tool's risk bins may be inconsistent across studies. A study of VRAG is of note here. Independent researchers compared recidivism rates in VRAG's nine scoring bins as represented in its experience table with the rates derived from multiple new studies.[157] Table 4 shows the differences between the original VRAG table with a selected three of the comparison studies.

[152] Timothy P. Cadigan et al., *The Re-Validation of the Federal Pretrial Services Risk Assessment (PTRA)*, 76(2) FED. PROB. 3, 4 (2012).

[153] *See generally* Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 ARIZ. ST. L. REV. 1 (2015), http://epubs.surrey.ac.uk/842340/1/47ArizStLJ1.pdf.

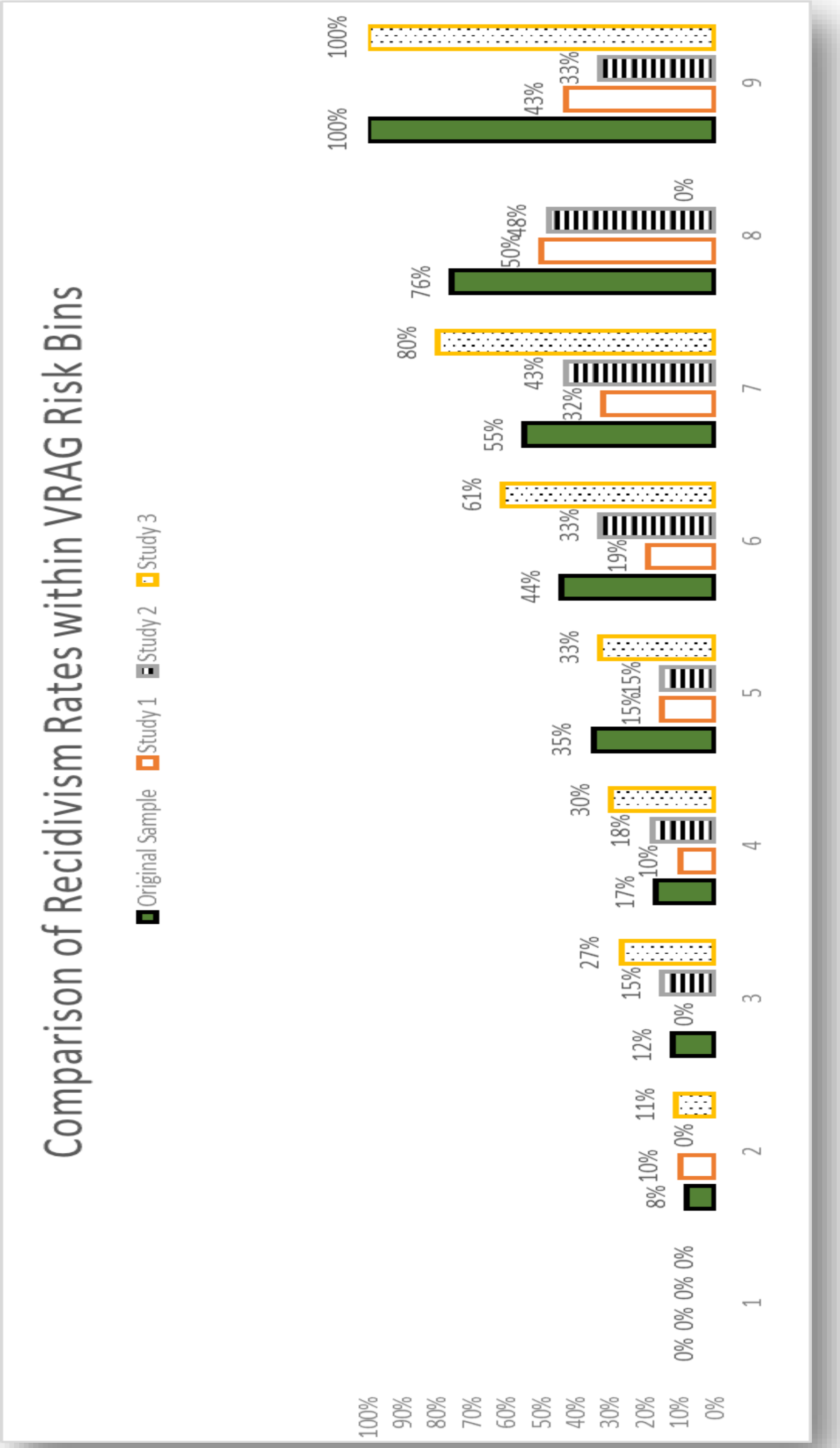
[154] Leslie Helmus et al., *Reporting Static-99 in Light of New Research on Recidivism Norms*, 21 THE FORUM 38 (2009), http://www.static99.org/pdfdocs/forum_article_feb2009.pdf.

[155] Leslie Helmus et al., *Reporting Static-99 in Light of New Research on Recidivism Norms*, 21 THE FORUM 38 (2009).

[156] *See generally* Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 ARIZ. ST. L. REV. 1 (2015).

[157] Astrid Rossegger et al., *Replicating the Violence Risk Appraisal Guide: A Total Forensic Cohort Study*, 9(3) PLOS ONE 1, 2 tbl. 1, 6 tbl. 4 (2014), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0091845.

Table 4: Comparison of Rates at VRAG Risk Bins



At each of the nine scores, the left-hand column represents the proportion in the original VRAG experience table. The three comparative studies agreed at score 1, with there being no recidivists. Yet notice the vast disagreement across the various studies with the other risk bins. Indeed, the idea of each score being associated with a monotonically higher rate of recidivism is not achieved in any of the three comparators.

There are other limits to validation. Typically, validation studies focus almost exclusively on how well the tool predicts failure (e.g., recidivism). While this may provide information on the risk prediction aspect, it does not sufficiently confirm how well the tool works with the partnered concerns of risk management or risk reduction.[158]

> *Policy Considerations:*
>
> *A tool must be cross-validated on the population and subpopulations on which it will be used, preferably before full implementation.*
>
> *Revalidation should occur at regular intervals to verify the tool's adequate performance and that the factors remain correlative with the outcome of interest.*

D. Reliability

Risk tools require human input in terms of the information needed by the algorithm. Because of such human involvement, tests for reliability are suitable. Reliability here means consistency in scoring by a single evaluator and across evaluators.

1. Inter-rater Reliability

One of the accuracy measures that should be regularly monitored addresses inter-rater reliability scores. These statistics reflect the degree of consistency of scoring across evaluators. Studies to date indicate wide-ranging reliability scores, indicating significant variability in consistency across tools and sites.[159] Some sites, though, achieved better scores than others using the same tool,[160] suggesting there may be mechanisms to improve inter-rater reliability statistics.

Assessments that require offender interviews may suffer if the evaluator does not possess appropriate interview skills and experience in reducing biases in responses. Further, interviewers should be trained in cultural sensitivities. To score items, inquiries may be required such as probing the individual's lack of empathy, amenability to treatment, prosocial personal relationships, lack of remorse, and attitudes toward authority.[161] It could be that offenders, particularly minorities, may appear uncooperative in interviews simply because of prior bad

[158] Stephane M. Shepherd & Danny Sullivan, *Covert and Implicit Influences on the Interpretation of Violence Risk Instruments*, 24 PSYCHIATRY, PSYCHOL. & L. 292, 295 (2017).

[159] Grant Duwe, *Why Inter-Rater Reliability Matters for Recidivism Risk Assessment* 2 (2017), https://psrac.bja.ojp.gov/ojpasset/Documents/PB-Interrater-Reliability.pdf.

[160] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 3, 20 (Jay P. Singh et al. eds., 2018).

[161] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 431 (2016).

experiences with authority figures, and this may also lead the assessor to score items differently as a result.[162]

> ℹ *Policy Considerations:*
>
> *Inter-rater reliability should be checked at regular intervals. Retraining may be necessary if reliability estimates are weak.*
>
> *If offender interviews are required to score a tool, evaluators should receive sufficient training in how to reduce interviewer bias and on culturally sensitive interview skills.*

2. *Overrides*

Risk assessment may not be entirely automated if a human remains in the loop with the ability to adjust risk results. Such adjustments are generally referred to as *overrides*. Two general types of overrides are known to occur. A policy override exists if agency officials believe that the risk tool's outcome should typically be overridden for a group of offenders whose unique risk factors do not appear to be sufficiently accounted for in the tool's factorology. Known examples are to adjust higher for sex offenders, young offenders with significant criminal histories, and offenders with severe mental health issues.[163] Policy overrides do not necessarily represent evidence-based practices. At times, jurisdictions employ particular policy overrides due to political sensitivities. The employment of a policy override for sex offenders, for instance, is often not based on evidence of underestimating risk of sex offending. Instead, such an override often is more about political fallout if sex offenders are not managed as high risk.

The second type is a professional override in which the evaluator believes there is something idiosyncratic about the individual case that is not sufficiently accounted for in the risk tool.

> One possible strategy is to employ a traditional risk/need assessment tool to obtain a baseline or "ballpark" risk level for the offender and then, depending on what idiosyncratic risk information might be available, augment, or override, the initial assessment by raising or lowering risk with the presence of additional risk and protective factors.[164]

For example, an evaluator may find the offender is in a state of acute stress, which in the evaluator's professional judgment is an imminent risk factor. The evaluator thereby may increase the prediction on the offender's likelihood of reoffending as a result. Alternatively, the evaluator may find that protective or promotive factors exist in the individual case that appear to lower the probability of reoffending.[165] In contrast, experts surmise that many professional overrides simply reflect the evaluator's personal distrust of the tool or an intent to manipulate it because

[162] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 431 (2016).

[163] *E.g.*, Thomas H. Cohen et al., *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80(1) FED. PROB. 12, 12 (2016).

[164] Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511, 1516 (2012).

[165] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 PSYCHOL. ASSESSMENT 1456 (2019).

of some preconceived judgment about the individual offender.[166] A professional override may also be used because of political sensitivities concerning certain offenders. Figure 7 reflects upon a probation officer (PO)'s experiences to illustrate this concept.[167]

Figure 7: Example of a Professional Override

> *"PO Balken supervises a man with mental health needs who often gets hostile as high risk. There is no reason he needs to be at a high risk level as there is nothing he can do with this probationer. But, in order to satisfy the public it must be on official records that he is supervised on high. If something were to happen and he killed his mother, the public would want to see that he was supervised on high supervision. It is basically for protection of the office. PO Balken argued if something were to happen while an individual was on low level supervision, the individual officer would not be protected because people do not trust the risk assessment."*
>
> Viglione (2019)

Override rates are imbalanced. Evidence to date indicates that the vast majority of overrides are toward higher risk rather than lower risk.[168] One reason may be the evaluators' attempts to reduce liability for potential false negatives.[169]

Overrides are not rare. Known override rates tend to be from 10% to 15% of cases,[170] though much higher rates have been observed at certain sites.[171] Likely, extreme or trivial override rates reflect institutional cultures in terms of trust in the risk tool used and the value placed on professional judgment calls.

[166] Jean-Pierre Guay & Geneviève Parent, *Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels with the LS/CMI*, 45 Crim. Just. & Behav. 82, 83-84 (2018).

[167] Jill Viglione, *The Risk-Need-Responsivity Model: How Do Probation Officers Implement the Principles of Effective Intervention*, 46 Crim. Just. & Behav. 655, 667 (2019).

[168] Angèle Christin et al., *Courts and Predictive Algorithms* 7 (Oct. 27, 2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf; Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 Crim. Just. & Behav. 1511, 1516 (2012).

[169] Katherine E. McCallum et al., *The Influence of Risk Assessment Instrument Scores on Evaluators' Risk Opinions and Sexual Offender Containment Recommendations*, 44 Crim. Just. & Behav. 1213, 1214 (2017).

[170] Thomas H. Cohen et al., *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80(1) Fed. Prob. 12, 15 tbl. 2 (2016).

[171] Fred Schmidt et al., *Predictive Validity of the Youth Level of Service/Case Management Inventory with Youth who have Committed Sexual and Non-Sexual Offenses: The Utility of Professional Override*, 43 Crim. Just. & Behav. 413, 421 (2016).

Political sensitivities are relevant to override rates as upward adjustments are far more likely with sex offenders than non–sex offenders across tools.[172] One study found a 34% rate of upward overrides for sex offenders compared with 14% for other types of offenders.[173]

While overrides are intended to allow for adjustments that the particular tool purportedly does not sufficiently address, evidence from the studies that exist show that overrides usually reduce the tool's accuracy overall.[174] For example, in a study of federal offenders, the recidivism rates of offenders whose supervision levels were discretionarily adjusted upward were equivalent to the rates of their initial algorithmic categories.[175]

Still, there is some evidence that overrides to a lower risk level may be appropriate. Researchers in one study found that sex offenders whose scores were adjusted downward recidivated less often and at rates consistent with their new, lower status.[176] The authors there surmised that it could be that the potential fallout for downward adjustments considering the popularity at the particular site of upward departures for sex offenders meant that these evaluators were quite strongly confident in their judgments.

An override policy may occur when something of importance has shifted environmentally in the jurisdiction after the tool was implemented and/or validated.[177] Potential possibilities include instances in which the base rates of reoffending have significantly changed or if the jurisdiction adopts successful interventions that will apply to the individual(s) being assessed to improve their chances of success. This type may be recognized as a policy override or as a case of professional judgment, depending on the circumstances.

E. Communicating Risk Tool Results

This section addresses issues with communication. The fact that tools were developed from studying historical data invites challenges to how to properly use *group* data to evaluate an *individual*'s future behavior. Algorithmic risk assessment is actually not an individualized task and thus is unable to convey an absolute prediction about the individual. It is therefore critical to understand that risk rankings (e.g., low versus high risk, or a greater-than-average risk) are relative to some group. Next, risk tools offer several options on how risk information is conveyed.

[172] Jean-Pierre Guay & Geneviève Parent, *Broken Legs, Clinical Overrides, and Recidivism Risk: An Analysis of Decisions to Adjust Risk Levels with the LS/CMI*, 45 CRIM. JUST. & BEHAV. 82, 91 (2018); Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511, 1520 (2012).

[173] Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511, 1525 (2012).

[174] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 PSYCHOL. ASSESSMENT 1456 (2019); Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511, 1523 (2012).

[175] Thomas H. Cohen et al., *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80(1) FED. PROB. 12, 12 (2016).

[176] Steven J. Wormith et al., *The Predictive Validity of a General Risk/Needs Assessment Inventory on Sexual Offender Recidivism and an Exploration of the Professional Override*, 39 CRIM. JUST. & BEHAV. 1511, 1530 (2012).

[177] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 PSYCHOL. ASSESSMENT 1456 (2019).

The selection matters given that research indicates that decision makers make judgments about an individual's risk that can vary depending on the communication type(s) the evaluator employs.

1. Group-to-Individual Challenge

Understanding the group-based nature of actuarial assessment tools is crucial. When attempting to determine the relative risk for an individual, the assessor's final score for the person is compared with the developmental data. The individual's risk level is ranked according to the frequency of recidivist acts observed in the development samples.[178] In other words, the tools were normed (i.e., standardized) on the developmental group studied.[179] As a result, group-based data, fundamentally, cannot absolutely provide information specifically attuned to the individual's risk.[180]

The reason for the potential mismatch is what has been nicknamed the "group-to-individual" or "G2i" challenge.[181] The *G* represents the discipline of science that studies a phenomenon at the group level; the *i* indicates that the law, conversely, seeks to use science to understand an individual.[182] The misapplication in an attempt to connect the two, the G2i path, is not entirely understood by legal practitioners. Therefore, law-oriented professionals often place too much emphasis on risk tool results in terms of adjudging the individual's level of risk. Group-based data can provide inferences about the group(s) from which it was derived but cannot diagnose any specific individual.[183] Thus, risk assessment here operates by deductive reasoning, as in Figure 8.

Figure 8: Deductive Reasoning

Risk assessment practices operate by deductive reasoning.

1. Those who scored a 6 were in the high-risk group.
2. Defendant Smith scored a 6.
3. Therefore, Defendant Smith is a high risk of reoffending.

In sum, the algorithmic score or category is not itself an individualized prediction. Where the tool offers an experience table, the relevant percentage is instead an estimate of the reoffending rate for the group that shares the score or category.

> [I]t is important to note that base rate data represent an average, with each person possessing a different propensity, and associated probability, to commit an act of

[178] John A. Fennel, *Punishment by Another Name: The Inherent Overreaching in Sexually Dangerous Person Commitments,* 35 NEW ENG. J. CRIM. & CIV. CONFINEMENT 37, 52 (2009).

[179] *See generally* Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 ARIZ. ST. L. REV. 1 (2015).

[180] Christopher Slobogin, *The Modern Case for Indeterminate Dispositions in Criminal Cases*, 48 SAN DIEGO L. REV. 1127, 1147 (2011).

[181] David L. Faigman, John Monahan, & Christopher Slobogin, *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 420 (2014).

[182] David L. Faigman, John Monahan, & Christopher Slobogin, *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 420 (2014).

[183] Brad Johnson, *Prophecy with Numbers: Prospective Punishment for Predictable Human Behaviour?*, 7 UTS L. REV. 117, 129 (2005).

> violence. Consequently, for some, the probability to engage in violence will be higher than the sample rate, and for others, the probability will be lower than the sample rate.[184]

Unfortunately, too many evaluators incorrectly place the group's risk estimate directly onto the individual. It is common to (erroneously) identify the recidivism rate of the group as an individualistic probability (e.g., "the score of 6 means that there is a 40% likelihood of Jones reoffending").[185] This type of attribution is incorrect and misleading. Deductive logic may be suitable to the medical sciences when there is evidence of causative factors to disease. The prediction of criminal behavior, though, is not based on causation. The predictive factors that are statistically significant, instead, are correlative. Hence, placement of a group estimate onto the individual is unjustified.

Another issue is that the individual is unlikely to be identical in risk-relevant ways to the developmental samples on which such estimates were derived. As a result, the accuracy of the group-based estimate will further suffer.

> The actuarial method compares similarities of an individual's profile to the combined knowledge of the past events of a convicted group of . . . offenders. An individual may share some, but typically not all, of the characteristics of the original sample. Hence, applying the results of an actuarial scale to an individual can have the effect of reducing the predictive accuracy of the scale. This is known as the "statistical fallacy effect."[186]

An alternative perspective is that these risk predictions resemble the concept of "naked statistics."[187]

Naked statistics refers to *"any information about a category of people or events not evidencing anything relevant in relation to any person or event individually. A piece of evidence is nakedly statistical when it applies to an individual case by affiliating that case to a general category of cases."*

Stein (2005)

[184] Michael H. Fogel, *Violence Risk Assessment Evaluation: Practices and Procedures, in* HANDBOOK OF VIOLENCE RISK ASSESSMENT AND TREATMENT: NEW APPROACHES FOR FORENSIC MENTAL HEALTH PROFESSIONALS 41, 47 (Joel T. Andrade ed., 2009).

[185] *See generally* Melissa Hamilton, *Adventures in Risk: Predicting Violent and Sexual Recidivism in Sentencing Law*, 47 ARIZ. ST. L. REV. 1 (2015) (providing examples from case law).

[186] Leam A. Craig & Anthony Beech, *Best Practice in Conducting Actuarial Risk Assessments with Adult Sexual Offenders,* 15 J. SEXUAL AGGRESSION 193, 203 (2009).

[187] ALEX STEIN, FOUNDATIONS OF EVIDENCE LAW 43 (2005).

Consider a scenario in which there are 25 workers, 24 of whom collectively murder their boss, while one of them takes no part in the homicide.[188] We shall call the latter Mr. Innocent. Prosecutors have clearly identified the 25 workers but cannot distinguish the 24 perpetrators from Mr. Innocent. Based on a group statistic that 24 out of 25, or 96%, of the group was involved, an algorithm would judge Mr. Innocent as having an extremely high probability (at 96%) of guilt. The way that the naked statistics work is to treat all 25 in the identified group the same—even, yes, Mr. Innocent. The algorithm has no way to distinguish which 24 of the 25 are guilty. Of course, this analogy is retrospective, rather than prospective, but the bones are the same regarding future risk.

Hence, algorithmic binning is not individualistic once the defendants are placed within the same score or categorical bin. In the criminal justice system's exaltation of the risk culture and the system's tendency to prefer false positives over false negatives, officials would likely move to preventively incapacitate in our murder scenario all 25 of them. They would justify this based on the 96% risk statistic, without much consternation for the unfairness to poor Mr. Innocent. In sum, our innocent worker is sacrificed to a statistical generalization. Still, the ramifications are felt individually by him.

Despite the poor fit between group statistics and individual predictions, criminal justice officials prefer pragmatism.

> [R]isk is a function of fitting a profile which is known to be statistically associated with causing a certain kind of harm and risk assessments are therefore necessarily based on probability data about classes of people which are insensitive to relevant but unknown differences among the individuals in that class. In seeking to prevent harm, the State will therefore be forced to rely on such data, notwithstanding their inability rationally to warrant inferences about particular individuals.[189]

A related complaint regarding the G2i challenge applies to criminal justice penalties based on risk: The person is not necessarily being sanctioned on his own merits. Penalizing a person via risk assessment derived from group data means that punishment becomes situated on shared group characteristics and thereby becomes too deindividualized.[190] The scheme is akin to punishing someone for what other (purported to be) statistically matched persons have done.[191] These notions of punishing someone for a group's misdeeds and the precrime (i.e., one that has not yet happened) will be fleshed out further below.

2. *Forms of Risk Communication*

Any specific tool may offer one or more methods of communicating the level of risk computed by the algorithm. Table 5 lists the most common.

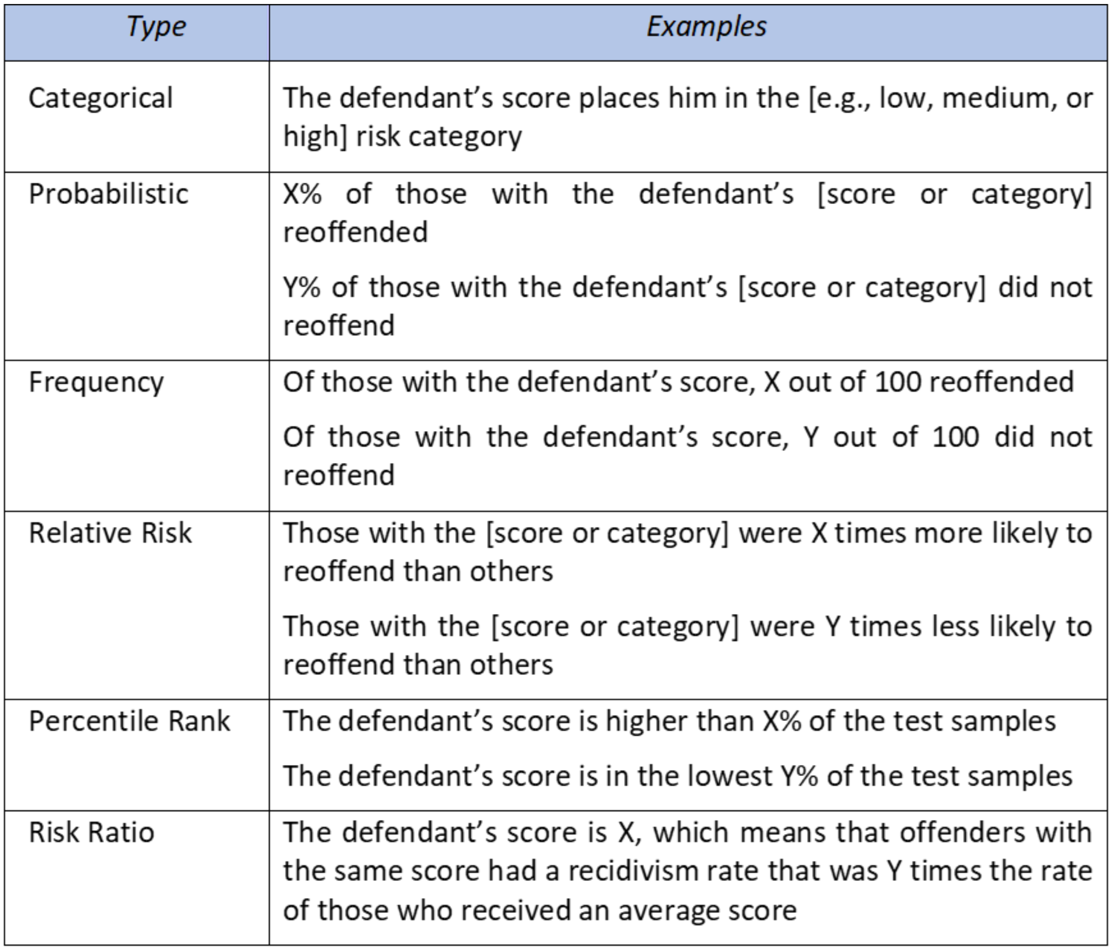
[188] The hypothetical is a modification of one presented earlier. Denise Meyerson, *Risks, Rights, Statistics and Compulsory Measures*, 31 SYDNEY L. REV. 507, 515 (2009) (in turn, crediting the paper's hypothetical to Charles R. Nesson, *Reasonable Doubt and Permissive Inferences: The Value of Complexity*, 82 HARV. L. REV. 1187, 1192-93 (1979)).

[189] Denise Meyerson, *Risks, Rights, Statistics and Compulsory Measures*, 31 SYDNEY L. REV. 507, 522 (2009).

[190] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 277 (2013).

[191] J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1390 (2011).

Table 5: Communication Strategies



Type	*Examples*
Categorical	The defendant's score places him in the [e.g., low, medium, or high] risk category
Probabilistic	X% of those with the defendant's [score or category] reoffended Y% of those with the defendant's [score or category] did not reoffend
Frequency	Of those with the defendant's score, X out of 100 reoffended Of those with the defendant's score, Y out of 100 did not reoffend
Relative Risk	Those with the [score or category] were X times more likely to reoffend than others Those with the [score or category] were Y times less likely to reoffend than others
Percentile Rank	The defendant's score is higher than X% of the test samples The defendant's score is in the lowest Y% of the test samples
Risk Ratio	The defendant's score is X, which means that offenders with the same score had a recidivism rate that was Y times the rate of those who received an average score

Categorical risk assessments are often the most desired form of communication.[192] They usually consist of statements that a person is of low, moderate, or high risk, or something similar. This type of risk communication, though, has been criticized for the following reasons:

> (i) lack of consistent definition and the number of categories across different risk assessment instruments; (ii) lack of consistency when experts interpret and communicate risk in legal proceedings; (iii) likelihood of leading decision-makers to overestimate rates of violence (e.g., by not providing base rate information for recidivism, which is lower than most decision-makers believe); and (iv) likelihood of being interpreted to have definitively answered the legal questions (i.e., a "high" risk offender must be "likely" to recidivate).[193]

[192] Ashley B. Batastini et al., *Does the Format of the Message Affect What Is Heard? A Two-Part Study on the Communication of Violence Risk Assessment Data*, 19 J. FORENSIC PSYCHOL. RES. & PRAC. 44, 46 (2019).

[193] Daniel A. Krauss et al., *Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical Versus Probabilistic Risk Communication in Sexually Violent Predator Decisions*, 36 BEHAV. SCI. & L. 532, 534 (2018).

Another objection is that dividing a population into a finite number of categories (e.g., low, medium, high) is too vague a method and unfortunately bundles together many individuals who actually are not very similar.[194] Because tools include many factors, it could be that any two individuals in the same risk bin share no common predictive factors; they simply share the same end score.

Studies indicate that algorithmic risk tool information may be misinterpreted without also providing the context of base rates. Lacking base rate information, decision makers tend to overpredict the probability of failure.[195] Such a tendency was confirmed in a recent study in which respondents, when given a categorical prediction of "medium risk" without numerical anchors such as base rates, presumed a higher likelihood of reoffending in percentage terms (mean of 60%) than was justified by the tool's developmental sample's actual outcomes at that level (17%).[196]

In another study, actual judges and forensic clinicians who were asked to estimate the sexual recidivism rate for a "high-risk" classification gave responses ranging from 5% to 100%.[197] Findings of several research projects substantiate the lack of consistent understanding among practitioners of what categorical risk bins mean in terms of the likelihood of reoffending. One study of forensic clinicians asked for a ceiling estimate for a "low"-risk individual: The mean response suggested that a 28% risk of recidivism should be the cutoff between low and moderate risk, but the range of a suggested cutoff threshold varied significantly between 8% and 54%.[198] The mean response for the cutoff between moderate and high risk in this same study was 69% with a range of 38% to 95%.[199]

Researchers elsewhere asked a sample of psychologists to estimate the likelihood of an offender labeled "high risk" to reoffend: The mean response was 64% with a standard deviation of 23%, meaning that two-thirds of the sample gave an estimate ranging from 31% to 87%, indicating a high degree of variability.[200] These various studies highlight discordance even among forensic professionals as to what categories such as low/medium/high actually signify in the real world.

[194] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 558 (2018).

[195] Ashley B. Batastini et al., *Does the Format of the Message Affect What Is Heard? A Two-Part Study on the Communication of Violence Risk Assessment Data*, 19 J. FORENSIC PSYCHOL. RES. & PRAC. 44, 56 tbl. 4 (2019).

[196] Ashley B. Batastini et al., *Communicating Violence Risk During Testimony: Do Different Formats Lead to Different Perceptions Among Jurors?,* 25 PSYCHOL. PUB. POL'Y & L. 92, 99 tbl. 3 (2019).

[197] Stephanie A. Evans & Karen L. Salekin, *Violence Risk Communication: What do Judges and Forensic Clinicians Prefer and Understand?*, 3 J. THREAT ASSESSMENT & MGMT. 143, 156 (2016).

[198] N. Zoe Hilton et al., *Does Using Nonnumerical Terms to Describe Risk Aid Violence Risk Communication?*, 23 J. INTERPERSONAL VIOLENCE 171, 179 (2008), https://www.researchgate.net/profile/Grant_Harris/publication/5686180_Does_Using_Nonnumerical_Terms_to_Describe_Risk_Aid_Violence_Risk_Communication_Clinician_Agreement_and_Decision_Making/links/004635239e225d2cd8000000/Does-Using-Nonnumerical-Terms-to-Describe-Risk-Aid-Violence-Risk-Communication-Clinician-Agreement-and-Decision-Making.pdf.

[199] N. Zoe Hilton et al., *Does Using Nonnumerical Terms to Describe Risk Aid Violence Risk Communication?*, 23 J. INTERPERSONAL VIOLENCE 171, 179 (2008).

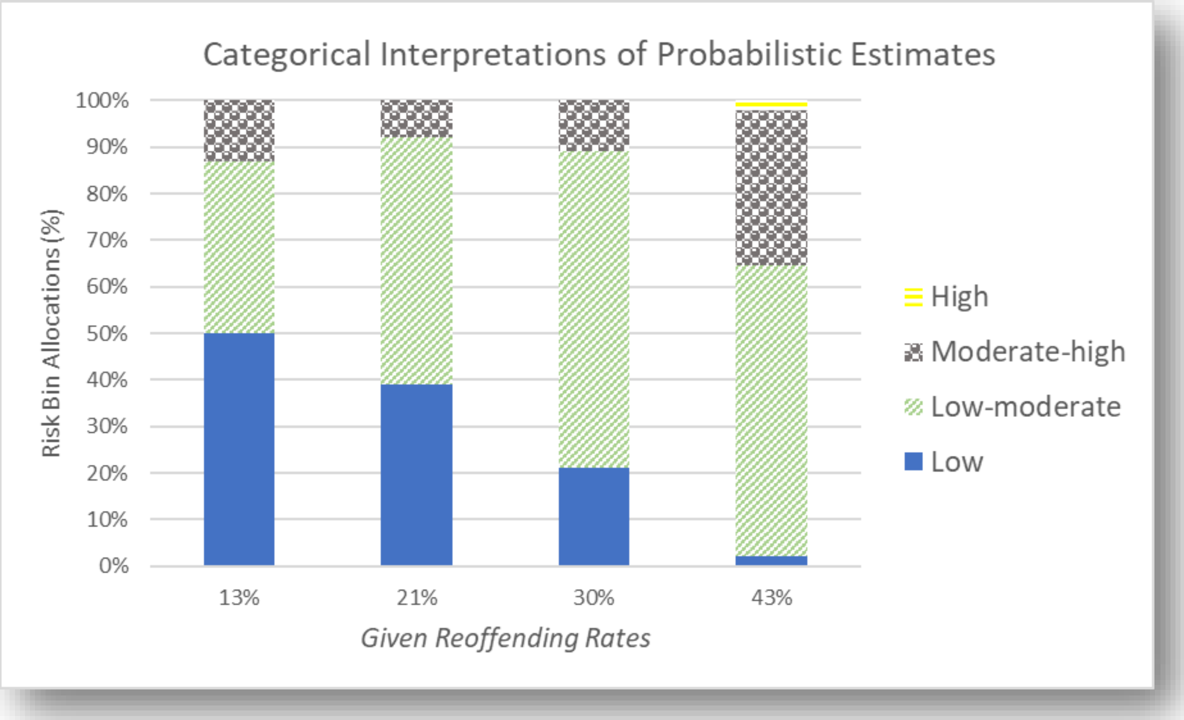
[200] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 556-57 (2018) (referring to relevant study).

Clearly, there is no consensus, even among practitioners in the field, as to what categorical bins mean, and they seem to have little concept of what probabilities signify.

> The existing empirical literature about risk categories suggests two conclusions. First, clinicians do not have a consensual understanding of the level of risk associated with a given risk category nor does the ascription of a particular risk category imply a particular level of risk. Since individuals with disparate risk levels will be given the same risk label (e.g., “high risk”), it is not surprising that there is a large variance in the rate of violence among individuals in the same risk category. Blunt risk categories obscure this variability by imparting the impression that individuals within the risk category are similar in terms of their risk of violence. However, this impression is not correct and therefore would hardly result in a better-informed decision.[201]

Not only do professionals have difficulty with translating categorical risk bins into appropriate percentages in terms of likelihood of reoffending, humans have problems in the other direction as well. Researchers in another study gave mock jurors probabilistic estimates of risk (e.g., 13%, 21%, 30%, and 43%) for a hypothetical sex offender.[202] The jurors were asked to place each probabilistic estimate into its appropriate categorical bin of low, low-moderate, moderate-high, or high. Figure 9 presents the results.

Figure 9: Survey of Recidivism Rates in Categorical Terms



[201] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 558 (2018).

[202] Daniel A. Krauss et al., *Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical Versus Probabilistic Risk Communication in Sexually Violent Predator Decisions*, 36 BEHAV. SCI. & L. 532, 541 tbl. 3 (2018).

Notice in Figure 9 the wide discrepancies in judgments as to where the specific probabilities fit into the categorical bins. At least some mock jurors thought that any of the four proportions could be judged in any of low, low-moderate, or moderate-high bins. Sizable portions of the mock jurors perceived a 13% likelihood of reoffending as low, low-moderate, or moderate-high. Groups of mock jurors also classified as moderate-high statistics ranging from 13% to 43%.

It is of note that evaluators can manipulate judgments on risk in the choice of communication strategies they use. In one study, mock jurors sampled were given alternative descriptions of a hypothetical sex offender who scored a 6. With the particular tool, a score of 6 was equivalent to a categorical "high-risk" ranking, a percentile ranking in which 31% of those with the same score sexually reoffended, and then a relative risk that those with a score of 6 had a recidivism rate 2.91 times that of the typical offender.[203] Despite these three being equivalent for the tool's score of 6, the jurors believed that the offender was a greater danger when the communication given was the categorical "high risk" than the other two variations.[204] Oddly, jurors in that same project portrayed two hypothetical offenders as equivalently risky whether they were told that their scores equated with a 9% versus a 31% recidivism rate.[205]

Decision makers in another study judged estimates using frequencies (one in 10 reoffended) as higher risk than when the same information is presented as equivalent percentages (10% of those in the group reoffended).[206] Possibly, the reason is that in the former scenario, it is easier to visualize at least one person who will commit a dangerous act.

Another study indicated that a different type of communication choice mattered. Mock decision makers were more severe when informed that there was a 26% likelihood of reoffending than when presented with the reciprocal that there was a 74% likelihood of not reoffending.[207] Obviously, these are identical statistics whereby one is simply the inverse of the other. This study suggests that a positive framing in terms of the risk of not offending could benefit offenders more than one that uses a negative framing about the prospect of failure.

3. Risk Rankings Are Relative

Risk bins often classify groups in an ordinal ranking and use categorical labels, though their numbers, names, and meanings vary. A critic decries the confusion that permeates risk assessment practices because of the lack of agreement about the meanings of groupings across

[203] Jorge G. Varela et al., *Same Score, Different Message: Perceptions of Offender Risk Depend on Static-99R Risk Communication Format*, 38 LAW & HUM. BEHAV. 418, 421-22 (2014).

[204] Jorge G. Varela et al., *Same Score, Different Message: Perceptions of Offender Risk Depend on Static-99R Risk Communication Format*, 38 LAW & HUM. BEHAV. 418, 421-22 (2014).

[205] Jorge G. Varela et al., *Same Score, Different Message: Perceptions of Offender Risk Depend on Static-99R Risk Communication Format*, 38 LAW & HUM. BEHAV. 418, 420-21 (2014).

[206] R. Barry Ruback et al., *Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis*, 80 FED. PROB. 47, 48 (2016).

[207] Neil Scurich & R. John, *Prescriptive Approaches to Communicating the Risk of Violence in Actuarial Risk Assessment*, 18 PSYCHOL. PUB. POL'Y & L. 50 (2011).

instruments.[208] Clinicians have no commonly agreed definition of risk categories,[209] statisticians have no accepted metric, and there are no normative or legal distinctions for such labels.[210] As a general rule, these categorizations are meaningless except as a rather crude ranking system. Individual risk tools may provide detailed context for each of its own categories (e.g., provide a definition and corresponding recidivism rate estimate), but these depictions will be unique to the specific tool. As an illustration, the recidivism rate in the developmental data for the "high-risk" group for one tool could be 80%, while in another tool 10%.

> There are no common meanings or understandings among clinicians, statisticians, evaluators, or decision makers of what low, medium, or high risk might mean in an absolute sense.

There are many reasons for inconsistency among risk tools on these issues with respect to binning and relative recidivism rates. The following are some examples:

- Recidivism rates for tools that define failure quite broadly (e.g., any supervision failure) will likely be much higher than when the definition is limited (e.g., conviction for a serious violent crime).
- Recidivism rates of bins with tools developed on data with higher base rates of reoffending will be higher than tools with lower base rate training samples.
- Some tool developers divide their training samples into equivalent groups (e.g., placing one-third into each of low, moderate, and high). For other tools, developers were less concerned with group size than distinguishing meaningful differences in rates. These choices will skew the number of individuals placed within bins.
- Risk binning based on a strategy of minimizing false negatives will present very differently than one in which the developers were keen to reduce false positives.
- Tools that engage in a binning strategy based just on risk factors will endow them with different meanings than tools whose bins are based on a combination of risk factors, needs, and/or protective factors.
- The follow-up period for the training data is relevant. A longer follow-up period will yield higher recidivism rates in bins simply because there is more opportunity to fail.

The categorical risk bin technique is merely a comparative and rhetorical device to differentiate the accumulation of risk factors among members of the tool's developmental sample. Depending on the predictors that are in the final algorithm and how they are weighted, individual offenders can receive inconsistent rankings across instruments. One particular study

[208] Daniel A. Krauss et al., *Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical Versus Probabilistic Risk Communication in Sexually Violent Predator Decisions*, 36 BEHAV. SCI. & L. 532, 534 (2018).

[209] Daniel J. Neller & Richard I. Frederick, *Classification Accuracy of Actuarial Risk Assessment Instruments*, 31 BEHAV. SCI. & L. 14 1, 142 (2013).

[210] J.C. Oleson et al., *Training to See Risk: Measuring the Accuracy of Clinical and Actuarial Risk Assessments Among Federal Probation Officers*, 75(2) FED. PROB. 52, 55 (2011).

highlights this concept. Researchers scored a sample of sex offenders using five standard violence and sexual recidivism actuarial tools and found disparate uses of high- and low-risk labels.[211] The authors of this study summarize their results as follows:

> [W]hen we attempted to identify sub-samples of high and low risk offenders using the [five] instruments, common sub-samples were not identified. An alarmingly high number (55% of the sample) were identified by at least one instrument as being high risk; an alarmingly small proportion of the sample (33% and 4%, respectively) was identified as either high or low risk by all [five] instruments.[212]

Similarly, another study of sexual recidivism tools found significant disagreement in the ordinal rankings of the same individuals who were assessed. The agreement between any two of the instruments in their low, moderate, and high designations ranged from 23% to 71%, but most agreed in fewer than half of the cases.[213] This means that a majority of individuals assessed would have landed in another risk bin if scored on an alternative tool.

In a similar vein, independent researchers compared findings from multiple studies that scored offenders on four violent recidivism risk tools.[214] Figure 10 provides a graphic of the annualized violent recidivism rates of offenders classified as high risk by either PCL-R, SORAG, Static-99, or VRAG.

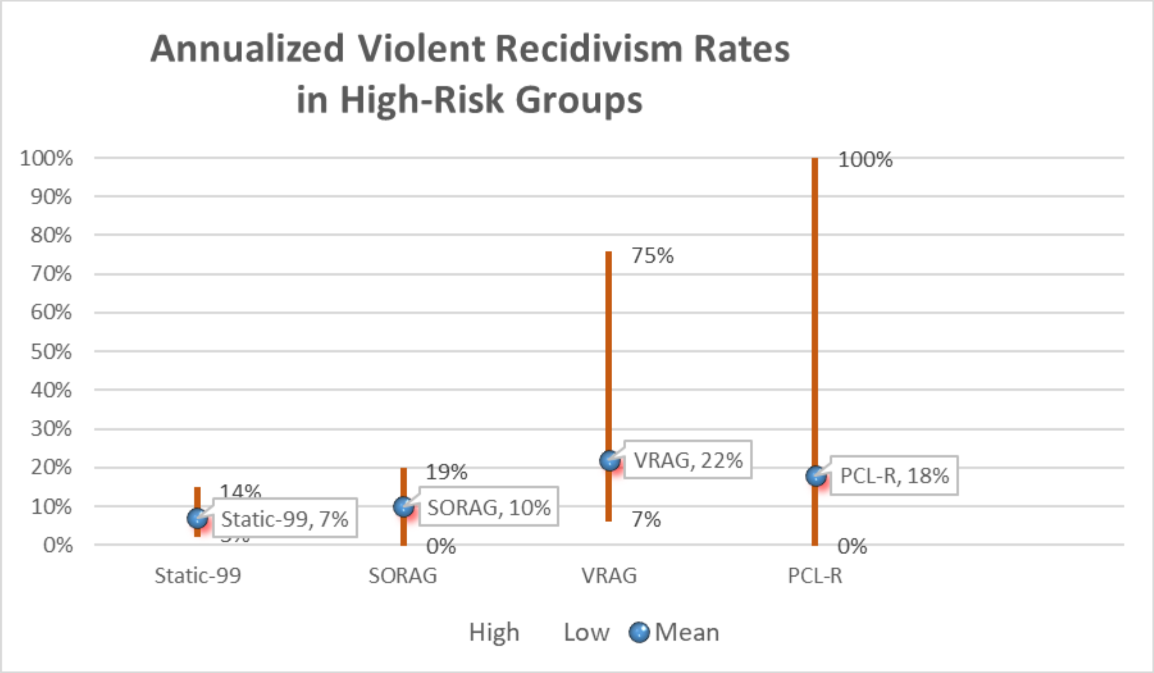
[211] Howard E. Barbaree et al., *Different Actuarial Risk Measures Produce Different Risk Rankings for Sexual Offenders*, 18 SEXUAL ABUSE 423, 429 (2006).

[212] *Id*. at 437.

[213] Sandy Jung et al., *Measuring the Disparity of Categorical Risk Among Various Sex Offender Risk Assessment Measures*, 24 J. FORENSIC PSYCHIATRY & PSYCHOL. 353, 361-62 (2013).

[214] Jay P. Singh et al*., Rates of Violence in Patients Classified as High Risk by Structured Risk Assessment Instruments*, 204 BRIT. J. PSYCHIATRY 180, 182 tbl. 1 (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3939440/.

Figure 10: Violent Recidivism Rates for the High Risk Across Tools and Studies



For each of the four risk tools, the two ends of the line indicate the lowest and highest recidivism rates from the relevant studies that scored each tool on a particular data set. The middle circle represents the mean recidivism rate for that tool's high-risk grouping across studies. Static-99 fares the best in having a fairly tight range of recidivism rates from 3% to 14% with a mean of 7%. PCL-R results are the most extreme, with high risk indicating recidivism rates from 0% to 100% depending on the study, and a mean of 18%. Across tools, the mean for high risk ranges from 7% to three times that at 22%. Not only is this a relatively wide range, but these rates also appear to be lower than many would expect for "high risk."

Thus, categorical labels have only relative meaning—not absolute value. They are comparators based on the attributes of the developmental samples on which the particular tool was based. To conceptualize what any category may mean, one must ascertain relevant details about the comparator group and context in which the tool was created.[215] For example, will the specific tool issue a label that compares the defendant to a population of adult males? Adult males who have previously committed a violent act? Violent adult males who are hospitalized in a forensic mental health setting? If the comparator group is of the latter (which is a real-life example, as VRAG was largely developed on such a sample), then the tool's relative rankings likely are inapplicable to an individual who is not similar to those in the normed group.

[215] Stephane M. Shepherd & Danny Sullivan, *Covert and Implicit Influences on the Interpretation of Violence Risk Instruments*, 24 PSYCHIATRY PSYCHOL. & L. 292, 297 (2017).

Risk assessment experts in the field have collaborated to craft a best-practices method for standardizing risk categories and their meanings.[216] In their white paper, these experts suggest a five-category ordinal ranking system, along with suggested percentiles of recidivists within each. However, there is little evidence that there is any momentum in the risk assessment field to widely adopt them.

Policy Considerations:

Agencies must make efforts to ensure that risk assessment communications are interpretable to the decision makers who receive them.

Communication of categorical rankings should be accompanied by appropriate base rate information relevant to the population to which the defendant belongs, with 95% confidence intervals.

Percentage estimates are preferred over relative risk as easier to understand. Still, 95% confidence intervals should also be offered.

Communications of the likelihood of succeeding (a positive framing) is preferable for many individuals.

The group-to-individual problem is important, and risk assessment outcomes based on group data cannot be placed onto individuals as if those outcomes were an absolute prediction.

[216] R. KARL HANSON ET AL., THE COUNCIL OF STATE GOVERNMENTS, A FIVE-LEVEL RISK AND NEEDS SYSTEM: MAXIMIZING ASSESSMENT RESULTS IN CORRECTIONS THROUGH THE DEVELOPMENT OF A COMMON LANGUAGE 3 (2017), https://csgjusticecenter.org/wp-content/uploads/2020/01/A-Five-Level-Risk-and-Needs-System_Report.pdf.

IX. Issues of Fairness and Bias

Recent controversies within the academic literature and publicized by news sources about algorithms in the criminal justice system orient toward concerns about whether they are fair, or alternatively, if they are biased. There is no agreed consensus on what the terms *fairness* or *bias* may mean in this context. In a scientific sense, bias often refers to a tool systematically underperforming in some way.[217] But even where a tool does not manifest systematic biases overall, it may still operate in ways that disadvantage individuals and/or groups. Bias may be perceived if the tool performs decently overall yet favors or disfavors a particular group in some fashion, whether intended or not, such as systematically erring to a greater degree with certain subpopulations.[218] Hence, bias within the algorithmic criminal justice world holds both statistical and social connotations.[219]

A. Systematic Bias Overall (i.e., Poor Accuracy)

A tool may be judged as too inaccurate as a general rule. As indicated earlier, one of the commonly cited reasons for the algorithmic turn is to reduce the negative impact of human biases on important criminal justice decisions. However, the algorithm itself has the ability to exacerbate bias. While human bias is acted out on a case-by-case basis, an algorithm's efficiency means it can discriminate on a more systematic basis and on a larger scale.[220]

A judgment about what evidence is indicative of overall inaccuracy could refer to a variety of measures, such as:

- an insignificant correlation coefficient between the tool's scores and recidivism (i.e., the scale and recidivism outcomes are unrelated)
- a low AUC for purposes of discriminatory validity
- an unacceptable high overall error rate
- an unacceptable calibration overall in that the *predicted* rates of reoffending in one or more scores or categories are significantly different (either higher or lower) than the *actual* recidivism rates in that same score or category
- an unacceptably high false positive rate, false negative rate, false discovery rate, and/or false omission rate

A tool that fails on one or more of these measures may be justifiably discontinued or deemed not worth the resources, unless officials see some other value, such as in its needs component. The more pressing problem, though, regards bias at the individual and/or group levels.

[217] Partnership on AI, Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System 15 (2019), https://www.partnershiponai.org/wp-content/uploads/2019/04/Report-on-Algorithmic-Risk-Assessment-Tools.pdf.

[218] Partnership on AI, Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System 15 (2019).

[219] Partnership on AI, Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System 15 n. 11 (2019).

[220] Indrė Žliobaitė, *Measuring Discrimination in Algorithmic Decision Making*, 31 Data Mining & Knowledge Discovery 1060, 1063 (2017), https://courses.helsinki.fi/sites/default/files/course-material/4595613/Zliobaite2017_Article_MeasuringDiscriminationInAlgor.pdf.

B. Measures of Fairness or Unfairness

No formal mechanism in the law or in the sciences exists to consistently enforce any form of algorithmic accountability in criminal justice.[221] This gap may explain how in real-world settings "algorithms (a) construct identity and reputation through (b) classification and risk assessment, creating the opportunity for (c) discrimination, normalization, and manipulation, without (d) adequate transparency, accountability, monitoring, or due process."[222]

1. *Individual Fairness*

Individual fairness requires that a tool's predictions be the same for similarly situated individuals. Here, that likely means that individuals (a) with the same predictive factors and (b) at the same levels (c) should receive the same risk score and (d) that such score would mean something similar in terms of recidivism outcomes. Individual bias would exist, for instance, if Persons A and B received a score of 6 but one was placed in low risk while the other in high risk.

2. *Algorithmic Group Fairness*

Interest in group fairness has emerged along with a new scientific literature on the topic called FATML, or "fairness, accountability, and transparency in machine learning."[223] The machine learning literature with FATML has produced a "staggering number of definitions of algorithmic fairness."[224] Table 6 contains some of the more common ones. Several of the measures within the table were introduced earlier.

[221] *See* Robyn Caplan et al., *Algorithmic Accountability: A Primer* 10 (Apr. 18, 2018), https://datasociety.net/pubs/alg_accountability.pdf.

[222] Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1218, 1239 (2017).

[223] Harsh Gupta, *Constitutional Perspectives on Machine Learning* 4 (Dec. 17, 2017), https://osf.io/preprints/socarxiv/9v8js/download?format=pdf.

[224] Philipp Hacker & Emil Wiedemann, *A Continuous Framework for Fairness* (Dec. 21, 2017), https://arxiv.org/pdf/1712.07924.

Table 6: Quantifying Group Fairness

Definition(s)	*Measure*	*Calculation*
Statistical parity Demographic parity	Acceptance Rate	$\frac{TP+FP}{N}$
Equalized odds Conditional procedural equality	True Positive Rate (TPR)	$\frac{TP}{TP+FN}$
Conditional procedural equality	True Negative Rate (TNR)	$\frac{TN}{TN+FP}$
False negative error rate balance Equal opportunity	False Negative Rate (FNR)	$1-TPR$
False positive error rate balance Equal opportunity	False Positive Rate (FPR)	$1-TNR$
Predictive parity Precision	Positive Predictive Value (PPV)	$\frac{TP}{TP+FP}$
Predictive parity	Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$
Conditional use error	False Discovery Rate (FDR)	$1-PPV$
Conditional use error	False Omission Rate (FOR)	$1-NPV$
Treatment equality	Cost Ratio of Errors	$\frac{FN}{FP}$ or $\frac{FP}{FN}$
Equal calibration	Predicted = Observed	$(TP+FP)=$ $(TP+FN)$

Various measures in Table 6 appraise different qualities of a tool. Further, several of them are mathematically inconsistent with one another.[225] The idea of total fairness, in which a tool would be deemed fair by all of the definitions in Table 6, is impossible in the real world.[226] This means that for any particular risk tool, one or more of the definitions may indicate the tool is compliant with group fairness, while at the same time exhibit group bias in one or more of the other measures. As a result, a critic will be able to assert a particular algorithm is biased simply by

[225] Thomas Miconi, *The Impossibility of "Fairness": A Generalized Impossibility Result for Decisions* 2 (Sept. 11, 2017), https://pdfs.semanticscholar.org/d883/b155d1ce19672cdf49795ea1a63acc923ad5.pdf.

[226] Richard Berk et al., *Fairness in Criminal Justice Settings: The State of the Art*, SOC. METHODS & RES. (forthcoming 2020).

selecting the measure of fairness that achieves this goal.[227] Conversely, a fan can select the measure that best exemplifies a tool's fairness.

To review these measures, we start in the top row. Statistical parity exists when the percentages of offenders predicted to recidivate and those predicted not to recidivate are the same across groups.[228] If, for example, 30% are predicted to recidivate in one group, the tool ought to predict 30% of the other group to recidivate. Statistical parity represents equal acceptance rates in that the tool is predicting (i.e., accepting) the same proportion of high risk across groups.[229] The literature also refers to this measure of equity as demographic parity if the groups at issue are distinguished by some demographic characteristic (e.g., race, class, gender).[230] The problem of differences in base rates is relevant here. If base rates vary, statistical parity cannot be met unless one adopts different thresholds by group. However, doing so would then reduce accuracy and violate individual fairness. If, for example, the base rate in Group A is 20% and in Group B is 40%, adjusting predictions to achieve the same prediction rate for both (e.g., 30%) to achieve statistical parity will necessarily undercut accuracy. Group A will simply have a lower-than-expected recidivism rate than Group B, and both groups will be misjudged, just in different directions.

Conditional procedural equality requires that the True Positive Rates (TPRs) and True Negative Rates (TNRs) be equivalent across groups. The group fairness idea of equal opportunity uses their reciprocals by requiring equivalent False Positive Rates (FPRs) and False Negative Rates (FNRs) across groups.

Moving down the table to the next two group fairness definitions, predictive parity envisions equivalent Positive Predictive Values (PPVs) and Negative Predictive Values (NPVs) across groups.[231] The reciprocals of these, the False Discovery Rates (FDRs) and False Omission Rates (FORs), represent conditional use errors in terms of being forecasting errors.

Table 7 provides another illustration from the Reporter's study of the COMPAS tool in Broward County regarding gender differences and how a tool may comply with some, but not all, of these measures. The purpose here is to show that, unlike in the previous examples concerning race,[232] two groups may instead be relatively compatible in discrimination measures (here using TPR and TNR) but yield different calibration metrics.

[227] Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC DAVIS L. REV. ONLINE 261, 289 (2019).

[228] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 184 (2019).

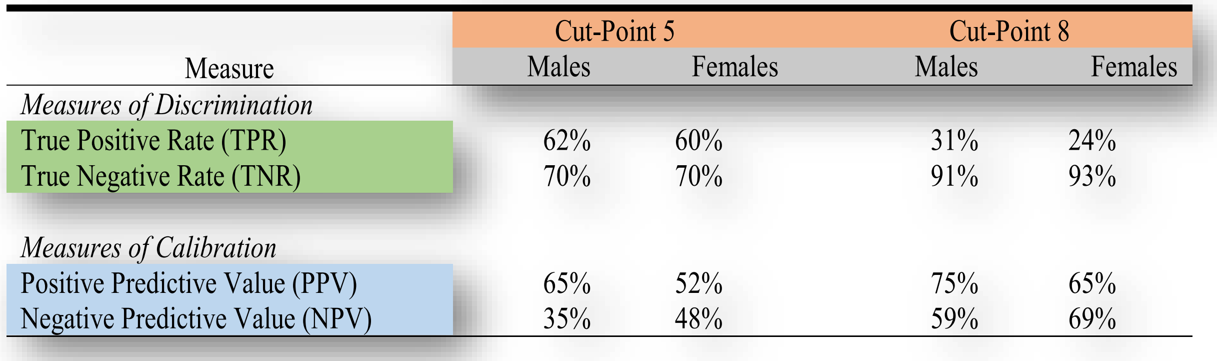
[229] Alexandria Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 1545 (2017).

[230] *See, e.g.*, James E. Johndrow & Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*, 13 ANNALS APPLIED STAT. 189 (2019), https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/30728?confirm=1d6331c2.

[231] Sahil Verma & Julia Rubin, *Fairness Definitions Explained* 4 (2018) (unpublished manuscript).

[232] *Supra* Table 2.

Table 7: Example of Disparities in Discrimination and Calibration



	Cut-Point 5		Cut-Point 8	
Measure	Males	Females	Males	Females
Measures of Discrimination				
True Positive Rate (TPR)	62%	60%	31%	24%
True Negative Rate (TNR)	70%	70%	91%	93%
Measures of Calibration				
Positive Predictive Value (PPV)	65%	52%	75%	65%
Negative Predictive Value (NPV)	35%	48%	59%	69%

Two different cut-points are offered. At the lower cut-point, notice relatively equivalent TPRs and TNRs between males and females. A proponent could claim here that COMPAS was unbiased toward genders using TPRs and TNRs as the empirical support. But a critic could point to the substantially unequal PPVs and NPVs to argue evidence of gender bias. For females, COMPAS is poorer at predicting recidivism and better at predicting non-recidivism.

The explanation for why the TPRs are relatively equal yet the PPVs vary significantly at least at the lower cut-point rests on the role of base rates. In the Broward County data set, the base rates of female versus male offender are substantially different. As a result of base rate differences, TPRs (generally immune to base rates) are similar (suggesting fairness), yet the PPVs (highly dependent on base rates) vary between them (signifying bias).[233]

Moving onto another measure, treatment equality considers the ratio of the errors, as in FN/FP or its reciprocal, FP/FN, and thus is also known as the cost ratio of errors.[234] Differences in how the tool prefers false positives over false negatives (or vice versa) between groups indicates group-based partiality. Using the Broward County data set and comparing whites versus blacks, the cost ratios of errors are unequal. The cost ratio of false positives to false negatives (FP/FN) is 0.7 for whites while it is 1.4 for blacks. Notice that the errors are in the opposite direction. This means that the COMPAS tool prefers false negatives for whites while contrastingly preferring false positives for blacks. In other words, the algorithm opts to err by wrongly classifying (recidivist) whites as low risk while wrongly classifying (non-recidivist) blacks as high risk.

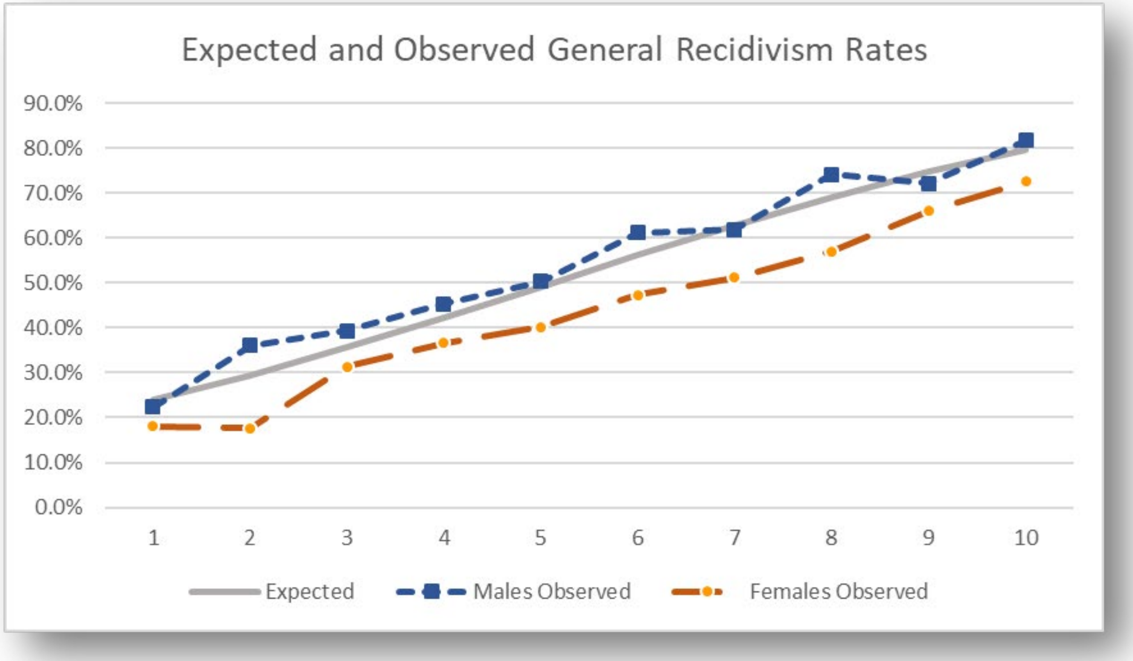
The final definition of group fairness in Table 6 considers equal calibration. The equal calibration metric in Table 6 is rather simplistic. Equal calibration is met if the number predicted to recidivate (expected) is equal to the number of recidivists (observed) and such equivalency is true across groups. Ideally, this definition would be met for a tool that predicted 30% of Groups A and B to reoffend and observed that 30% of Groups A and B did reoffend. Still, the expected versus observed recidivism rates can be better evaluated by breaking these comparisons down across a tool's ranking system. Using the Broward County data set again to illustrate for

[233] Melissa Hamilton, *The Biased Algorithm*, 56 AM. CRIM. L. REV. 1553, 1574-75 (2019).

[234] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 181 (2019), https://onlinelibrary.wiley.com/doi/pdf/10.1111/jels.12206.

differences by gender, Figure 11 plots the expected recidivism rate for the groups combined[235] across COMPAS's scoring system from 1 to 10 (with higher numbers predicting greater risk of reoffending).

Figure 11: Example of Lack of Equal Calibration by Gender

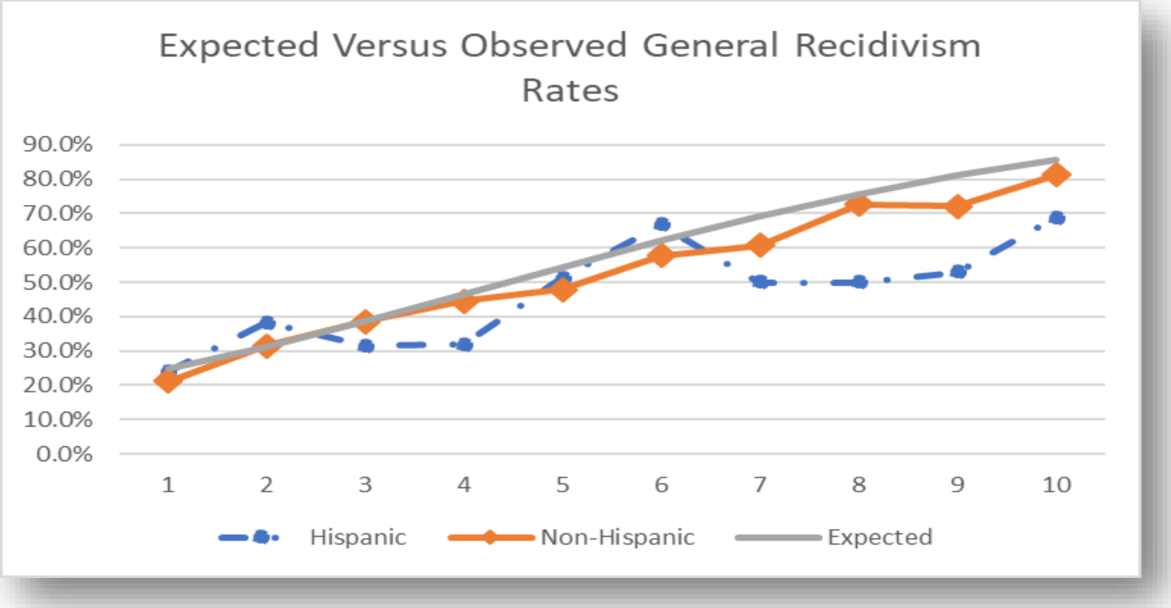


The straight line represents the expected (i.e., predicted) rate of recidivism by decile score for the genders combined. The upper dotted line represents the actual recidivism rates of males at each decile score. Notice the observed rates of reoffending for males track relatively closely to the straight (expected rate) line. This tracking indicates that COMPAS performs decently (is calibrated well) in predicting recidivism for males. However, the lower dotted line for females demonstrates that COMPAS does a decent job at scaling for females but systematically overpredicts risk for women at all deciles. Females consistently reoffended at lower rates than COMPAS predicted.

The situation is more complicated for Hispanic and non-Hispanic groups. Figure 12 and Figure 13 use the same Broward County data set to show the pattern for general recidivism and violent recidivism, respectively.

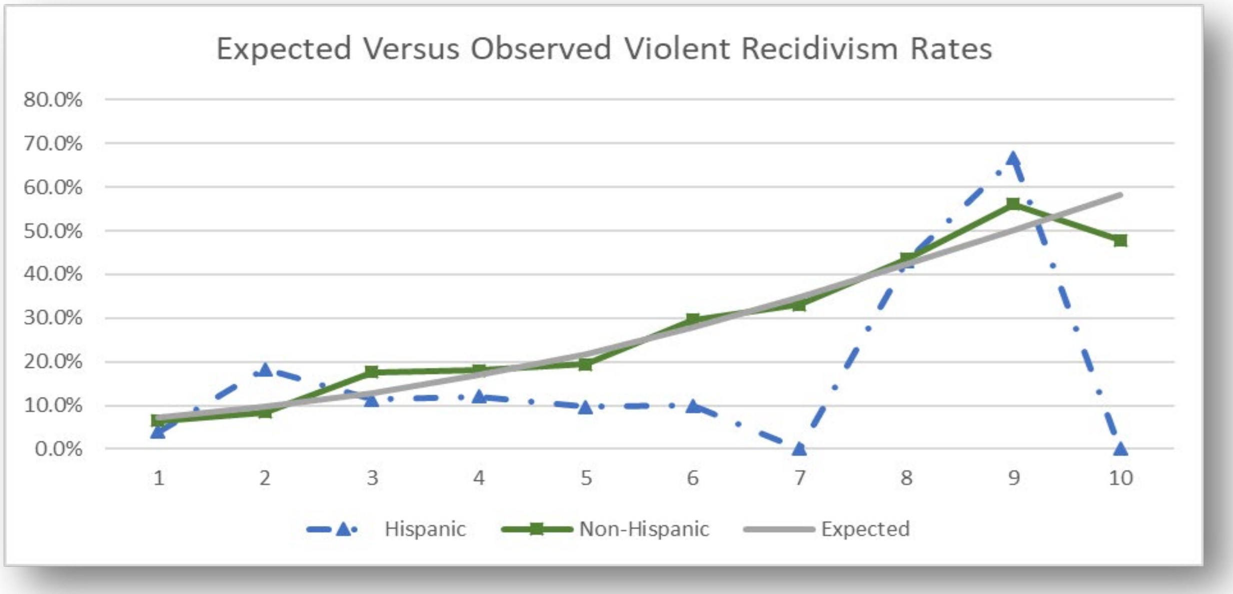
[235] This was accomplished using a logistic regression equation with decile score as the predictor and general recidivism as the dependent variable.

Figure 12: Example of Lack of Equal Calibration by Ethnicity for General Recidivism



Notice that the non-Hispanic rates of recidivism track very closely to the overall expected rate line. Yet the Hispanic rates significantly vary from the expected rate line. Indeed, one can observe that COMPAS decile scores have a curvilinear relationship with rates of general recidivism. The Hispanic data points indicate failures in both discrimination and calibration. The relationship is even worse with the violent recidivism scale as shown in Figure 13.

Figure 13: Example of Lack of Equal Calibration by Ethnicity for Violent Recidivism



For the COMPAS violent recidivism scale, the more extreme curvilinear relationship for Hispanics indicates performances on discrimination and calibration are quite poor.

Alternative metrics for algorithmic fairness were not presented in Table 6. A special calibration measure called "balance for the positive class" requires that the mean test score for those in the positive class (i.e., recidivists) are equivalent across groups.[236] Correspondingly, "balance for the negative class" requires equivalent mean test scores for those in the negative class (i.e., non-recidivists). For example, an algorithm would be considered unbiased if the mean score for Group A recidivists was 4 points while the mean score for Group B recidivists was also 4 points.

An illustration from the COMPAS tool with the Broward County data set using race may be useful in depicting a violation of these class balancing calibration metrics in Table 8. Note that here the table shows only a dichotomization of white or black and thus excludes other races/ethnicities.

Table 8: Example of Imbalanced Calibration for Race

	General Recidivism Scale	
	Recidivists	Non-Recidivists
Black	6.2	4.2
White	4.7	2.9

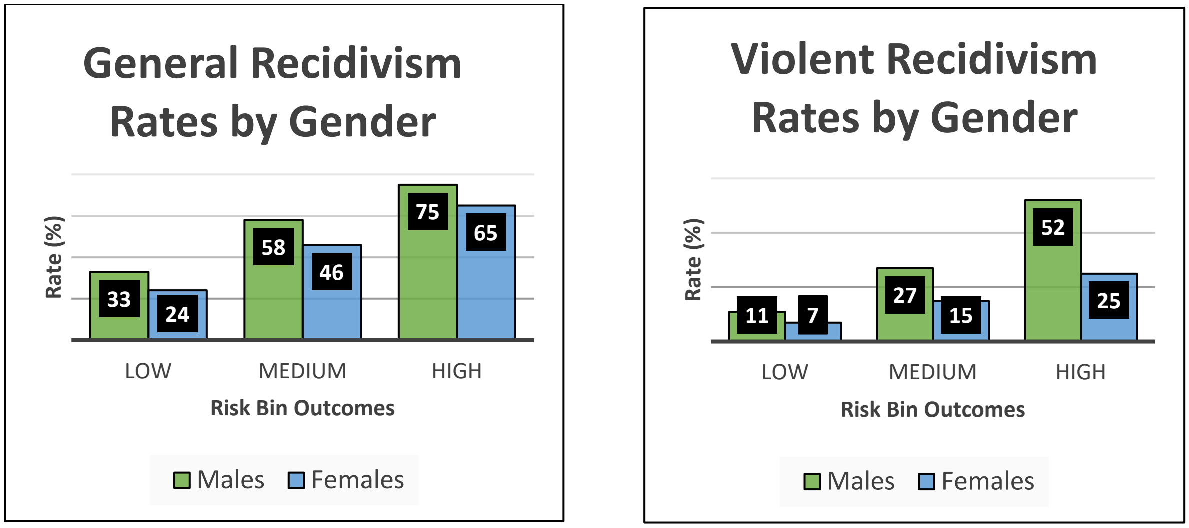
One can see in Table 8 that in the general recidivism scale, the mean score (out of a scale from 1 to 10) of recidivists for blacks was 6.2 while for whites it was 4.7. The mean score for non-recidivists for blacks was 4.2 while it was 2.9 for whites. These results violate the balance for the

[236] Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Conf. on Innovations in Theoretical Computer Science 2 (2017), https://arxiv.org/abs/1609.05807.

positive and negative classes. Scores on COMPAS simply don't mean the same for each group.

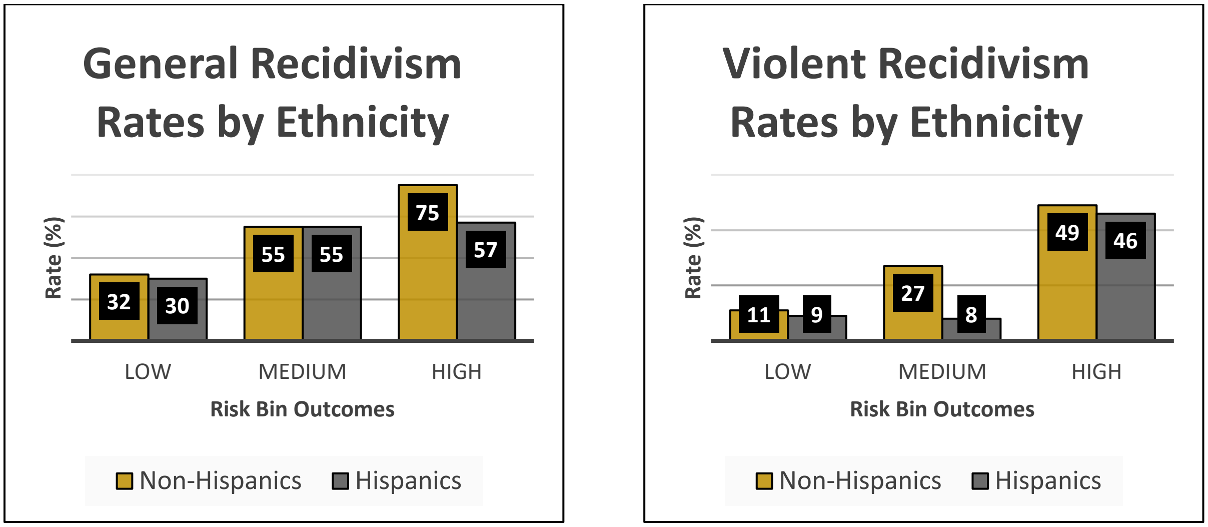
The foregoing ways of analyzing group fairness are not the only ones available. Another method is rather simplistic and looks at whether categorical bins are accurate and equal between groups. Figure 14 uses the Broward County data set again and shows the recidivism rates by gender within the COMPAS-scored categorical risk bins using both its general and violent recidivism scales.

Figure 14: Example of Risk Bins and Observed Recidivism by Gender



Notice that males reoffended at greater levels than females for both general and violent recidivism. Indeed, at the extreme in the high-risk violence tool, females violently reoffended at half the rate of males. Figure 15 does the same using Hispanic ethnicity with the COMPAS general and violent recidivism scales.

Figure 15: Example of Risk Bins and Observed Recidivism by Ethnicity



Hispanics recidivated at lower rates at each risk bin and for each of the general and violent recidivism scales with one exception: The medium-risk bin for general recidivism reflects equal failure rates. Also of note is that the three-risk-bin strategy does not work well with Hispanics. For general recidivism, there is only a 2% differential between the medium- and high-risk bins for Hispanics. Then in the violent risk scale, the Hispanic violent recidivism rate was slightly lower in the medium bin compared with the low-risk bin. It appears a two-risk-bin strategy is best for Hispanics, though with a different binning strategy for each tool. For general recidivism it appears that combining the medium and high groups would work while for violent recidivism combining the low and medium groups fits the data better.

COMPAS is not the only tool whose risk bins show disparate recidivism rates across groups. The federal Post Conviction Risk Assessment (PCRA) tool performs disparately based on age groupings, according to a study.[237] Using age categories of 25 and younger, 26 to 40, and over 40, the PCRA showed that the youngest group recidivated at a greater rate and the oldest group at a lesser rate comparatively across most of PCRA's categorical bins.[238]

3. Test Bias

Another statistical model exists for analyzing group bias, but it has not yet permeated the risk assessment literature beyond a select few studies.[239] The gold standard for investigating test bias is endorsed by the American Psychological Association and was honed by evaluating academic tests in educational contexts. Test bias here refers to a systematic error in how a test measures members of one group as compared with another group.[240] This methodology is imminently appropriate for the algorithmic risk assessment world in determining whether a tool has an equal relationship with the outcome for both groups. A statistical explanation for this test bias methodology is beyond the scope of this report, but further information is easily accessible for the interested reader.[241]

Studies using this test bias model show varied results. For example, in analyses of the COMPAS data set in Broward County, COMPAS systematically overpredicts general recidivism for females.[242] The federal Post Conviction Risk Assessment instrument likewise indicates test bias against women.[243] In comparison, research on the federal Pretrial Risk Assessment tool found test bias in a pretrial release context of overpredicting for women in violent arrests but no test bias for the tool involving any arrests.[244]

[237] John Monahan et al., *Age, Risk Assessment, and Sanctioning*, 41 LAW & HUM. BEHAV. 191 (2017).

[238] John Monahan et al., *Age, Risk Assessment, and Sanctioning*, 41 LAW & HUM. BEHAV. 191, 199 fig. 2 (2017).

[239] Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC DAVIS L. REV. ONLINE 261, 291-92 (2019) (citing studies).

[240] Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 738 (2009).

[241] *See generally* Melissa Hamilton, *The Biased Algorithm*, 56 AM. CRIM. L. REV. 1553 (2019).

[242] Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAV. SCI. & L. 145, 151 (2019).

[243] Jennifer Skeem et al., *Gender, Risk Assessment, and Sanctioning*, 40 LAW & HUM. BEHAV. 580, 585 (2016).

[244] Thomas A. Cohen & Christopher Lowenkamp, *Revalidation of the Federal Pretrial Risk Assessment Instrument (PTRA): Testing the PTRA for Predictive Biases*, 46 CRIM. JUST. & BEHAV. 234, 253 tbl. 8 (2019).

Test bias was also shown by COMPAS against Hispanics on both general and violent recidivism scales.[245] In contrast, researchers studying PCRA found it was not biased against blacks even though a greater percentage of higher-risk predictions were given to blacks (largely because of criminal history).[246] Finally, statisticians focused on COMPAS represented that it did not exhibit test bias for blacks.[247]

C. How Biases May Enter Algorithms

Algorithms are not, as some might assume, neutral and impersonal in character. Algorithms are innately value-laden. "Operational parameters are specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others."[248] Several avenues are available for bias to enter, as outlined herein.

1. Label Bias

Developers choose their desired outcome variable (i.e., how to define the failure event) and its two options (e.g., recidivist versus non-recidivist).[249] As examined earlier, issues exist with the problematic reliance on certain behaviors as constituting recidivism. Technical violations, supervision revocation, institutional misconduct, and disciplinary problems are weak substitutes for criminal offending. They are even less justifiable as risk tools should generally be more focused on *serious* offending. Many of these tools as a result will produce biased results for users who (quite reasonably but erroneously) assume that these algorithms are isolating to serious behaviors that also constitute crimes. What these tools are forecasting then may be biased estimators of what users expect to be predicted.

In any event, when the outcome of interest involves *crime*, the measurement of this failure event is innately biased.[250] There is simply no practical or theoretical way to measure crime *per se*. Thus, developers must resort to using proxies. By definition, though, any proxy for crime will be fundamentally inaccurate.[251] Developers vary in what proxies they use for their crime variable. Common proxies are technical violations, supervision failures, arrests, convictions, or reincarcerations. But these may be more representative of official *responses* to crime than of offenders' behaviors.[252] As none of these are synonymous with actual crimes committed, the resulting inaccuracies create noise in the algorithm.

[245] Melissa Hamilton, *The Biased Algorithm*, 56 AM. CRIM. L. REV. 1553, 1570 (2019).

[246] Jennifer L. Skeem & Christopher T. Lowenkamp, R*isk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 700 (2016).

[247] Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses*, 80(2) FED. PROB. 38, 43 tbls. 5, 6 (2016) (however, to get to this result researchers failed to comply with the gold standard protocol by including additional factors [e.g., gender and age]).

[248] Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, BIG DATA & SOC'Y 1, 1 (July-Dec. 2016), https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679.

[249] Maddalena Favaretto et al., *Big Data and Discrimination: Perils, Promises and Solutions: A Systemic Review*, 6(12) J. BIG DATA 1, 12 (2019), https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4.

[250] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 18 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.

[251] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 18 (2019).

[252] PARTNERSHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE SYSTEM 17 (2019), https://www.partnershiponai.org/wp-content/uploads/2019/04/Report-on-Algorithmic-Risk-Assessment-

Proxies for crime convey multiple types of overlapping errors and gaps. For various reasons, not all crimes are known. Victims may not identify what was done as crimes in the first place. For additional reasons, many crimes are never reported. Even if reported, police may not respond or take a report. Certainly not all crimes result in a record of arrest, and fewer still ever result in convictions. Attrition rates suggest that victim reports and arrests are questionable markers of actual crime statistics.

Thus, even when tool developers rely on official records, results are already biased. When using official records, developers tend to prefer arrest records because of their better availability and because they require shorter follow-up periods than convictions. Yet arrests are problematic for various reasons. An arrest is "one of the least procedurally protected instances of contact with the criminal justice system."[253] The low evidentiary bar as a result renders arrest records highly unreliable and problematic. Arrest data suffer from discretionary actions, even discriminatory motivations, by police. "Officers use discretion in enforcement decisions (e.g., deciding whom to stop, search, question, and arrest) just as police officers and prosecutors use discretion in charging (e.g., simple assault vs. felonious assault). The underlying data reflect… these judgment calls."[254] The training data may thereby learn on what amounts to overpolicing practices in minority neighborhoods and underpolicing in upper-class areas. As a result, the algorithm may overestimate the risk of minorities while underestimating the risk of offending by whites.[255]

Arrest records may in any event be factually inaccurate in portraying whether an individual committed a crime.

> Police may arrest the wrong suspect or may arrest for behavior that turns out not to be criminal at all once a full investigation has been completed. Charges may be brought against the wrong defendant or may not align with the actual behavior in which the defendant engaged. The fact that many of these cases do not proceed to conviction gives rise to doubt about whether a crime occurred at all or whether an error was made by system actors themselves.[256]

The length of the follow-up period used to study recidivism in the testing sample will likewise skew results. Some crimes are more readily detected and the responsible parties identified quickly. For example, a short follow-up period is more likely to pick up street-level drug use than white collar fraud schemes,[257] resulting in a biased tool toward both types of offenders.

Tools.pdf; Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[253] Jessica M. Eaglin*, Constructing Recidivism Risk*, 67 EMORY L.J. 59, 94 (2017).

[254] EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS 22 (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[255] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 18 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.

[256] Cecelia Klingele, *Measuring Change: From Rates of Recidivism to Markers of Desistance*, 109 J. CRIM. L. & CRIMINOLOGY 769, 786-87 (2019) (internal citations omitted).

[257] Cassie Deskus, *Fifth Amendment Limitations on Criminal Algorithmic Decision-Making*, 21 LEGIS. & PUB. POL'Y 237, 248 (2018).

At times, proxies for crime are multilayered. For example, VRAG includes as its main attribute of violent reoffending any arrest for a violent crime. In order to include institutionalized offenders in their training data, VRAG developers chose to also count any violent acts that, in the opinion of agency employees, would have resulted in arrests if committed outside the institution.[258] Hence, VRAG uses a sort of proxy for a proxy in its outcome variable of violent recidivism.

While not technically a label bias problem, the fact that developers focus on failure rather than success contorts the whole scheme. Instead, an algorithm that learned on data to predict desistence or successful reentry would look very different yet still would likely be useful for the same decision points that risk assessments inform. Perhaps the myopic laser on failure filters through criminal justice officials' tendency to fear the false negative over the false positive.

2. *Feature Selection*

Bias may be embedded within the collection of predictive factors scored by the algorithm. Crime is a complex issue. Faced with multifaceted, real-world scenarios in which crimes occur, developers must select from a multitude of possible predictors. Despite how many dozens of factors are scored, an algorithm cannot possibly present a complete picture of the circumstances in which crimes are perpetrated. The choice of predictors, known as feature selection, is thereby an inherently reductionist exercise.[259] Simply put, every tool oversimplifies crime.[260] The point is that developers introduce bias by choosing which factors to test in the first place and then narrowing to a smaller number of predictors to incorporate into their final algorithms.[261]

Factors that are selected (and those that are omitted) may reflect explicit or implicit bias on the part of the developers. Failing to include, or minimizing the weights of, predictive factors that are more culturally sensitive to females and minorities exemplifies feature bias as well. The lack of diversity in the data scientists working with algorithms (the vast majority are white males) augments potential avenues for bias to go unrecognized and thus uncontrolled.[262]

3. *Specification Error*

Developers make judgment calls about how to define and measure the selected predictors in ways that may introduce specification error. Let us use VRAG as an example. VRAG was designed to predict violence. Choices were required regarding which types of acts would and would not count as violence. VRAG developers acknowledged that certain acts of child molestation might not be violent per se. But these developers intentionally opted, because of their own value judgments on the subject, to deem all criminal acts of sexual contact as constituting violent acts, even in the absence of threat or aggressive behavior (e.g., statutory rape).[263] Yet they further

[258] VERNON LEWIS QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK 123 (1998).

[259] Maddalena Favaretto et al., *Big Data and Discrimination: Perils, Promises and Solutions: A Systemic Review*, 6(12) J. BIG DATA 1, 14 (2019), https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4.

[260] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 18 (2019), https://tlsprdsitecore.azureedge.net/-/media/files/topics/research/algorithms-in-criminal-justice-system-report-2019.pdf?rev=ffc06e85e9c244ceaa9f160f27a8b2b3&hash=D1F64FAFF4FBE536DA22B6599C10E5D9.

[261] Maddalena Favaretto et al., *Big Data and Discrimination: Perils, Promises and Solutions: A Systemic Review*, 6(12) J. BIG DATA 1, 14 (2019), https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4.

[262] Nuria Oliver, *The Tyranny of Data, in* ASSESSING THE IMPACT OF MACHINE LEARNING ON BEHAVIOUR 58, 61 (Emilia Gómez ed., 2018).

[263] VERNON LEWIS QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK 122 (1998).

declared that violence would not include the sexual crimes of possession of child pornography, exhibitionism, and voyeurism.[264] Notice these discretionary choices that might normatively conflict with others' assumptions on the matter (including those of users).

Making a different value-laden decision, the developers of HCR-20—designed to predict violence—boldly broadened their definition of violence to include threats of serious psychological harm. These examples from VRAG and HCR-20 also overlap with label bias.

Hence, the way factors are defined and scored may introduce other sorts of bias. As another example, a variable on employment stability may misrepresent the predictive nature of it for many women. Women may take time out of the workforce because of parental or family responsibilities, which usually are more protective behaviors that reduce recidivism risk. Thus, the model might be weakened if the algorithm is not nuanced enough to adjust for how certain predictive factors, specified in a way that may be suitable for the majority, may be erroneously applied to a subgroup.

Many tools score criminal history regardless of how old that history is. Failure to adequately address the passage of time introduces error as studies indicate that the predictive nature of criminal history erodes substantially over time.[265] Additional issues with respect to criminal history are discussed further below.[266]

A predictive model will exemplify specification error when factors that correlate with recidivism are excluded, a phenomenon also referred to as omitted variable bias.[267] One reason for omitted variable bias is the nature of sampling in the test data. Training samples typically are limited to offenders who were released from custody.[268] This means that the algorithms may not be learning on individuals who were not released, which may represent the more dangerous offenders as implied by that very fact.

Developers may intentionally exclude statistically relevant factors because of a pragmatic awareness that evaluators will have a difficult time finding the evidence to efficiently and properly code them.[269] Indeed, second generation tools (still in use today) are intended to be easily and quickly scored using a rather small amount of readily available data. To offer such efficiency, developers might purposely omit relevant variables, reducing model fit as a result.

Omitted variable bias occurs, too, when an instrument excludes a variable that is correlated with both an existing predictor and the outcome. Figure 16 is an example of this sort of omitted variable bias using the VRAG again. A VRAG factor concerns elementary school maladjustment. It could well be that low self-control is an explanatory factor for both elementary school

[264] VERNON LEWIS QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK 122 (1998).

[265] Melissa Hamilton, *Back to the Future: The Influence of Criminal History of Risk Assessments*, 20 BERKELEY J. CRIM. L. 76, 124-25 (2015) (citing studies), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3171520_code827096.pdf?abstractid=2555878.

[266] *See infra* Section VIII.B.4.

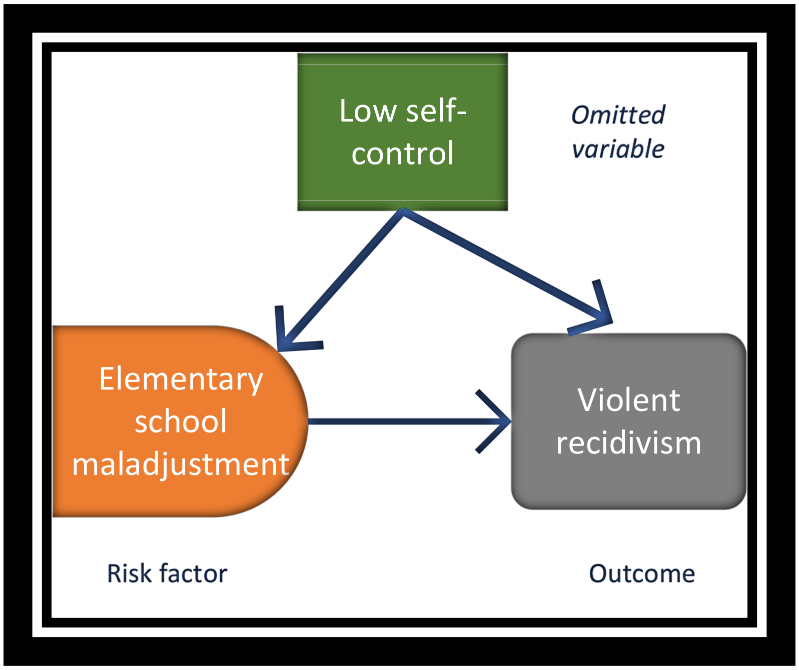
[267] Saul Levmore & Frank Fagan, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. (forthcoming 2020).

[268] Saul Levmore & Frank Fagan, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. (forthcoming 2020).

[269] Mariana Valverde et al., *Legal Knowledges of Risk, in* LAW AND RISK 86, 102 (Law Comm'n of Canada ed., 2005).

maladjustment and violent recidivism. As VRAG does not score self-control, this would then exemplify omitted variable bias.

Figure 16: Example of Omitted Variable Bias



There is a natural limitation to predictive factors as well. Algorithms can process only items that reasonably can be defined and quantified.[270] Could one sufficiently define and calculate, for instance, self-efficacy? Plus, there are certain potentially *causative* factors that are not accessible or easily amenable to scoring (e.g., potential roles of genetic predisposition or brain damage to violent acts).

The type and degree of specification error may also depend on the statistical method chosen to create the algorithm. The available options vary dramatically within the risk assessment literature. Modeling has ranged from the following: choosing predictors and weights from a literature review; simple correlations; regressions; decision trees; and supervised machine learning. Even an unsupervised machine learning program design has been theorized, but no tool available today appears to be a result of a completely human-free design process. These methodological choices will certainly result in different algorithms and carry unique sets of biases as a result.

4. Sample Bias

Ideally, the training data would be sufficiently representative of the population on which the tool will be used in a real-world setting.[271] A failure of representativeness exemplifies sample bias.

[270] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 18 (2019).

[271] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 19 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.

Unlike the best-practices standard in empirical research, risk tool development typically does not rely on independent, random samples. Instead, test samples are most often dependent samples (e.g., prisoners in the same institution or being treated in the same program) and draw on convenience samples (the data happened to be accessible).[272] Selection bias may result here wherein the samples were available because the officials in charge of the relevant data are simply more transparent. This scenario could make such a jurisdiction unique because it suggests the local officials are more culturally progressive in ways that permeate their system and thereby make their data sets less generalizable.

The developmental process may be statistically flawed because certain groups are under- or overrepresented in the training data.[273] The result is that knowledge about the outcome of interest will be stronger for the overrepresented group but less certain for the underrepresented one.[274]

Limitations in the convenience samples chosen to date only increase the potential for bias along the lines of race/ethnicity and gender. Many of the tools used in the United States were trained on data in other countries, usually Canada and European countries with predominately white populations.[275] Even in domestically developed tools, the samples are often not very diverse. These limitations in sampling mean that minorities and women tend to be underrepresented in developmental samples for most risk assessment tools.[276] In sum, algorithms tend to be normed on largely white male samples and are biased as a result.[277] Risk factors therein thus are often more salient to white male offenders, while meaningful risk factors that are culturally relevant to minority or female populations may be omitted. A risk assessment process that presumes that risk tools are somehow universal, generic, or culturally neutral will result in misestimation, as expressed in the following quote.[278]

[272] Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1096-97 (2013).

[273] Maddalena Favaretto et al., *Big Data and Discrimination: Perils, Promises and Solutions: A Systemic Review*, 6(12) J. BIG DATA 1, 13 (2019).

[274] Harsh Gupta, *Constitutional Perspectives on Machine Learning* 3 (Dec. 17, 2017), https://osf.io/preprints/socarxiv/9v8js/download?format=pdf.

[275] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 3, 4 (Jay P. Singh et al. eds., 2018).

[276] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity*, 22 PSYCHOL. PUB. POL'Y & L. 427, 428 (2016).

[277] Kelly Hannah-Moffat, *Algorithmic Risk Governance*, 23 THEORETICAL CRIMINOLOGY 453, 458 (2019).

[278] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 429 (2016).

"The over or under estimation of risk that can ensue from this process is entirely plausible given (a) the potential omission of meaningful risk items specific to minority populations, (b) the inclusion of risk factors that are more relevant to White offenders, and (c) variation in the cross-cultural manifestation and expression of existing risk items."
Shepherd & Lewis-Fernandez (2016)

It is improper to assume algorithms perform as well for culturally diverse groups because of risk-relevant differences.

> While instruments may be thought of as "ethnically neutral," "universal," or "generic," this view may overlook genuine cross-cultural differences in behavioral practices and expectations, health beliefs, social/environmental experiences, phenomenology, illness narratives, deviant conduct, and worldview.[279]

It is not surprising, when risk tools are originally normed on largely white samples, that at least some studies show risk tools provide more accurate predictions for whites than other groups.[280] With respect specifically to Hispanic Americans, academics have suggested that risk assessment tools may not perform well if they fail to "consider the centrality of family, acculturation strain, religiosity, gender role expectations, and culturally stoic responses to adversity" unique to this particular cultural group.[281]

Relatively few revalidation studies on ethnic minorities exist.[282] A few recent validation studies that have included an ethnic minority group may be telling. The data reported with a revalidation study of the federal Pretrial Risk Assessment tool showed that the tool overpredicted for Hispanics on any pretrial arrest (but not on violent pretrial arrest), while underpredicting blacks on any violent pretrial arrest (but not any pretrial arrest).[283] A recent study of the COMPAS tool indicates a significant degree of overprediction for Hispanics compared with non-Hispanics.[284] A meta-analysis of the Level of Service Inventory (LSI) family of tools found that

[279] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 Psychol. Pub. Pol'y & L. 427, 429 (2016).

[280] Jay P. Singh et al., *Comparative Study of Violence Risk Assessment Tools: A Systematic Review and Metaregression Analysis of 68 Studies Involving 25,980 Participants*, Clinical Psychol. Rev. (2011); Jay P. Singh & Seena Fazel, *Forensic Risk Assessment: A Metareview*, 37 Crim. Just. & Behav. 965, 978 (2010).

[281] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 Psychol. Pub. Pol'y & L. 427, 428 (2016).

[282] T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 Eur. Psychiatry 134, 135 (2017), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408162.

[283] Thomas A. Cohen & Christopher Lowenkamp, *Revalidation of the Federal Pretrial Risk Assessment Instrument (PTRA): Testing the PTRA for Predictive Biases*, 46 Crim. Just. & Behav. 234, 247 tbl. 4, 250 tbl. 6 (2019).

[284] *See generally* Melissa Hamilton, *The Biased Algorithm*, 56 Am. Crim. L. Rev. 1553 (2019).

while the scales tended to rate minorities at higher risk, the scales' predictive abilities were lower with minorities.[285]

Regarding gender, an algorithm may produce biased scores for women if it fails to incorporate gender-sensitive attributes relevant to female offending. Women's offending is impacted to a greater extent by parental stress, personal relationship problems, prior victimization, and effects of trauma.[286] A systematic review of cross-validations of nine risk instruments revealed significant variability in performance for women across studies and tools, which the authors attribute to the failure to fully consider a gendered perspective regarding the onset and maintenance of criminal careers.[287]

Some questions in current instruments are implicit proxies for gender. For example, the COMPAS tool queries whether the individual believes they have the requisite skills to obtain a minimum-wage job. Yet employability and wage gaps suggest that women will disproportionately answer in the negative. It perhaps is not surprising then that a recent study of the COMPAS tool, in a jurisdiction that chose not to use the available gender-specific scales, showed that it systematically overpredicted risk for women at all levels.[288]

5. Biases Preexisting in Training Data

Risk algorithms learn on historical data. This training data may incorporate information reflecting (and reifying) preexisting discriminatory decisions. A significant source of bias derives from a disproportionately heavy reliance on criminal history factors.[289] Clearly, criminal history information in the training data may represent discriminatory arrest practices by police based on sociodemographic characteristics (e.g., racial/ethnic affiliation, immigration status, gender). Prosecutorial policies that impose a disproportionate burden on minorities may also introduce bias into the training data. For example, initiatives such as no-drop policies to deter a particular problem (e.g., gun violence, street-level drug dealing) may increase conviction rates for minorities to a greater degree than nonminorities. If unchecked, the resulting algorithms would thereby learn that such sociodemographic traits (or proxies thereof) are predictive of offending.

Bias may be exacerbated when the training data are produced by the same actors as those who will use those predictive tools.[290] Consider a situation in which police target a certain neighborhood and the arrest rates of area residents thereby increase. This arrest data is then used to inform an algorithm. In turn, the algorithm predicts higher risk of recidivism for neighborhood residents, leading to more rearrests, and thereby entrenching overpolicing

[285] Mark E. Olver et al., *Thirty Years of Research on the Level of Service Scales: A Meta-Analytic Examination of Predictive Accuracy and Sources of Variability*, 26 PSYCHOL. ASSESSMENT 156, 168-69 (2014).

[286] Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAV. SCI. & L. 145, 147 (2019).

[287] Kate Anya Geraghty & Jessica Woodhams, *The Predictive Validity of Risk Assessment Tools for Female Offenders: A Systematic Review*, 21 AGGRESSION & VIOLENT BEHAV. 25, 32 (2015).

[288] *See generally* Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAV. SCI. & L. 145 (2019).

[289] *See generally* Melissa Hamilton, *Back to the Future: The Influence of Criminal History on Risk Assessment*, 20 BERKELEY J. CRIM. L. 75 (2015).

[290] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 18 (2019).

practices. The scheme becomes circular. The algorithm's prediction of an individual's being subject to overpolicing is itself predicted by a past history of people like him being overpoliced.[291]

Another source of misestimation in arrest data is that police officers are not generally trained in data collection techniques. Information recorded in police records may be erroneous simply because of a lack of training, lack of consistency in recording, and potentially few incentives to ensure the accuracy of the data. We can use race and ethnicity as exemplars here. If one officer typically relies on the suspects' self-reported race or ethnicity, those records may systematically differ from a fellow officer who relies on his/her own judgment calls. These sorts of input errors are not necessarily the result of malfeasance or negligence on the part of individual officers. Race and ethnicity do not have standardized meanings. There is simply no metric or norm for ascribing how much African genetic heritage is required to assign the label *black*. Ethnicity is even more difficult. Hispanic status is considered an ethnicity, rather than a race, such that one can choose to adopt the Hispanic culture. Without a detailed coding sheet and training on how to score race/ethnicity factors, they are open to uninformed, biased, or prejudiced representations.

One wonders as well how much missing data in the training sets are unknown because the recording officers simply guessed in order to fill in requisite blanks. If, for example, the suspect has no identification but the police report requires a birthday or age, it must be possible that at least some officers simply fill in what might seem to be plausible numbers. The developers may also introduce bias in attempting to clean the data by making decisions for how to correct for apparent errors and/or impute missing data.[292] Then there exists the possibility of what has been referred to as "masking," whereby developers intentionally manipulate the data to improve accuracy statistics that as a result may skew against one or more particular groups.[293]

Policing or prosecutorial practices are not the only sources of bias in the training data. Other sociodemographic characteristics (e.g., educational attainment, housing stability, steady employment) that implicate prejudicial actions by other actors (e.g., teachers, landlords, employers) may also introduce bias. The training data may also reflect broader, systemic bias representing societal inequities, such as living in a high-crime area, long-term unemployment, and untreated mental health problems.

6. Conflicts of Interest

Algorithmic biases may be perpetuated because of incentives developers have—consciously or not—to mask the biases. An allegiance effect occurs when the developers' belief in the superiority of their own tool inhibits impartiality. Examples of allegiance bias include poor ability to objectively evaluate the accuracy of one's tool, failure to reveal limitations with respect to whether the tool performs equally across groups, or suppressing other information about potential biases within or created by their algorithm.

Observers suggest that conflicts of interest occur because the purported success of their algorithmic tools may attract more users, financial rewards, professional acclaim, and/or career

[291] Discussion at Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[292] Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 80 (2017).

[293] Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L. J. 2218, 2260 (2019).

opportunities.[294] Studies find some evidence of allegiance effect. Compared with independent researchers, tool developers tend to report higher levels of predictive ability for their own tools.[295] This is also referred to as authorship effect.[296] The existence of authorship effect plagues even tools whose developers have been more transparent in making relevant information available. For instance, reviewers evaluating the performances of VRAG and Static-99 had these comments:

> Across 48 effects examined in the current meta-analysis, the validity coefficients were larger when conducted by instrument authors than when conducted by others. Although effects tended to be larger in initial validation studies than cross-validation studies, effects from the instrument authors' cross-validation studies were larger than those from nonauthors' cross-validation studies. These findings suggest that there is, indeed, a pattern of allegiance in the actuarial risk assessment literature, at least when allegiance is defined as being an author of an instrument.[297]

Tool developers naturally have conflicts of interest with their own algorithms.

It is of note that tool developers specifically in the field of criminal justice risk assessment act contrary to typical industry standards of empirical research in that they generally fail to admit to potential conflicts of interest in their research publications.[298] It is troubling that the current state of the criminal justice algorithmic risk field seems to condone the fact that most of the validation and cross-validation studies are performed by developers and others with financial and professional incentives tied to promoting positive performance statistics.

[294] Jay P. Singh et al., *Authorship Bias in Violence Risk Assessment?*, 8 Plos One 1, 2 (2013).

[295] Howard N. Garb & James M. Wood, *Methodological Advances in Statistical Prediction*, 31 Psychol. Assessment 1456 (2019); Pamela R. Blair et al., *Is There an Allegiance Effect for Assessment Instruments?: Actuarial Risk Assessment as an Exemplar*, 15 Clinical Psychol. 346, 346 (2008).

[296] Jay P. Singh et al., *Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis*, 8 Plos One 1, 2 (2013).

[297] Pamela R. Blair et al., *Is There an Allegiance Effect for Assessment Instruments?: Actuarial Risk Assessment as an Exemplar*, 15 Clinical Psychol. 346, 354 (2008), https://www.academia.edu/18230972/Is_There_an_Allegiance_Effect_for_Assessment_Instruments_Actuarial_Risk_Assessment_as_an_Exemplar.

[298] Seena Fazel, *The Scientific Validity of Current Approaches to Violent and Criminal Risk Assessment, in* Predictive Sentencing: Normative and Empirical Perspectives 197, 202 (Jan W. de Keijser et al. eds. 2019); T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 Eur. Psychiatry 134, 135 (2017).

Policy Considerations:

Attention to potential conflicts of interest is required in the development, modification, and validation of algorithmic tools. Cross-validation studies should be outsourced to independent researchers.

Stakeholder involvement in the development of the algorithm is recommended so that they are aware of the potential for bias to enter algorithms and thereby provide relevant guidance. Policy choices will be required because of potential tradeoffs between bias, accuracy, individual fairness, and group fairness.

7. Feedback Loop

Once embedded, biases can become further entrenched. Algorithms may suffer from a feedback loop in which biases are amplified over time. Biased predictions create additional inequalities from which the algorithm learns and then skews future predictions even more.[299] As an example, if a sentencing jurisdiction uses a biased algorithm, assigning higher risk predictions may mean that minorities receive more punitive punishments and thus are seen as more dangerous, thereby magnifying sentencing disparities in the future.

D. Efforts to Remediate Bias

Can anything be done about bias? The response to the potential for biases should not be simply to remove offending factors or ignore group differences. Redacting factors or group-sensitive attributes that statistically correlate with recidivism necessarily reduces the algorithm's predictive abilities, thereby undermining any benefits to criminal justice decision-making. Considering that the tools may already be less accurate for minorities and women (as evidenced earlier), removing statistically significant factors may further erode their viability with those groups. Decreasing accuracy of the tools for minorities and women would thereby further disserve them because of resulting increases in error rates applying to them. It is also problematic that by excluding variables because they correlate with a protected group, societal discrimination and inequalities are thereby concealed.[300]

Data scientists are actively hypothesizing various ways to ameliorate bias while attempting to maintain some minimal validity levels, whether it be preprocessing, in-processing, or post-processing.[301] Nonetheless, these options are largely theoretical; little is known about whether tool developers or users are employing any of them in real-world applications. Further, any remediation is likely to require a trade-off with some other values.

In a preprocessing option, to adjust for larger error rates for minorities and women, developers could oversample those groups to improve their ability to recognize culturally specific risk factors to incorporate. This option would require that developers test different sets of factors

[299] JOSHUA NEW & DANIEL CASTRO, CTR. DATA INNOVATION, HOW POLICYMAKERS CAN FOSTER ALGORITHMIC ACCOUNTABILITY 5 (2018), https://www.datainnovation.org/2018/05/how-policymakers-can-foster-algorithmic-accountability.

[300] Flavio P. Calmon et al., *Optimized Data Pre-Processing for Discrimination Prevention* 1 (Apr. 11, 2017), https://arxiv.org/pdf/1704.03354.

[301] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 184 (2019).

with these groups. In other words, the scientists would have to intentionally use a lens that focuses on the salience of race/ethnicity and gender. Such an exercise may appear to scientists as discriminatory. But a better view is one of cultural sensitivity and awareness that predictors of human behavior can vary across groups of people. Another preprocessing option is to clean variables that correlate with a protected category such that the new, cleaned variables are no longer associated with the protected category.[302] There is a cost to accuracy as a result because cleaned variables mean that relevant information has been excised from the data set.[303]

If base rates differ between groups, one could provide weights to equalize them.[304] Hypothetically, if the base rate of criminal history for Group A is three times that of Group B in the training data, the algorithm could discount the criminal history for Group A by a factor of three. This option would serve to better equalize the groups on criminal history yet maintain rank order within groups.[305] If groups vary on a sociodemographic variable because of race-based societal discrimination (e.g., educational attainment, employability), the algorithm could weight those factors less heavily for minorities. Yet these options prefer group fairness over individual fairness because individuals with the same education or employment record would be treated differently. It is not clear what this type of remediation would do for accuracy.

It might also be possible to reduce the potential for biased patterns in variables measuring crime by comparing rates across data on arrests, convictions, victimization surveys, and self-report surveys.[306] This effort might shed light on how biased the crime-related variables were in the training data and allow for some relevant adjustment in the algorithm.

One could manipulate output scores between groups to equalize error rates. This could include applying different scales and/or threshold cut-points. Still, if base rates vary, equalizing one error rate will increase one or more other error rates. Moreover, equalizing error rates might require treating the groups differently in that a single score means different things for the groups.[307] Equalizing error rates would violate the algorithmic fairness definition of equal calibration.

Developers could create entirely different algorithms trained on specific groups.[308] Multiple algorithms could weight factors differentially and/or include unique factors within each. For example, an algorithmic tool designed for Hispanics might include risk and protective factors that are ethnically sensitive to them. Or the protected group status (e.g., race, gender) could be

[302] Daniel McNamara et al., *Trade-offs in Algorithmic Risk Assessment* (Aug. 31, 2018) (unpublished manuscript), http://www.ong-home.my/papers/mcnamara18domestic-violence.pdf.

[303] Daniel McNamara et al., *Trade-offs in Algorithmic Risk Assessment* (Aug. 31, 2018) (unpublished manuscript).

[304] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 186 (2019).

[305] Rachel K.E. Bellamy, *AI Awareness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias* (Oct. 3, 2018), https://arxiv.org/pdf/1810.01943.

[306] PARTNERSHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE System (2019).

[307] Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L. J. 2218, 2275 (2019).

[308] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 18 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.

explicitly incorporated as a predictive factor.[309] Several tools offer gender-sensitive versions.[310] Legal and ethical issues with these culturally sensitive options are further addressed in Section IV.

Overall, it is not surprising that algorithms contain such biases. Algorithms used with human beings as subjects are inherently *discriminating* devices—they are designed to differentiate among individuals and groups based on observable characteristics. In turn, observable traits tend to be sociodemographic in nature. The tools thereby serve to either *privilege* or *disparately impact* certain groups.

The likelihood of choosing an inappropriate algorithm may be amplified when it is common that employees tasked with purchasing a tool are not sophisticated in these issues, may not demand a validation study on the population for which the tool is expected to be used, and/or may not order an independent analysis of the fit of the tool to the particular context.[311]

Policy Considerations:

Stakeholder involvement in the remediation of bias is essential for helping to ensure that the methods are appropriate for the jurisdiction. Stakeholders should be alert to biases built into risk assessment tools, both at the inception of the tool and when they are working with tools that have already been created.

X. Legal and Ethical Issues

A foundation has been laid for the law and ethics related to risk assessment. A judge has put defense counsel on notice that awareness of legal issues with respect to algorithmic risk assessment is expected for adequate representation as the following quote evidences.[312]

> "[A] reasonably competent attorney should be aware of potential avenues of attack on risk assessment tools that are well established in the legal literature. A quick computer-based search related to risk assessments would draw an attorney into the vibrant literature and developing caselaw related to use of risk assessment."
>
> *State v. Guise* (2018)

The judge may overstate the "well-established" nature of existing case law, but the basics are present. This section reviews a variety of these issues (known and anticipated) within the field.

[309] Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 18 (Aug. 14, 2018) (unpublished manuscript).

[310] *See infra* text accompanying notes 320-325.

[311] Discussion at Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[312] State v. Guise, 921 N.W.2d 26, 33 (2018) (Appel, J. concurring).

A. Risk Factors

Risk tools inform criminal justice decisions that differentiate between various groups with respect to benefits or burdens and may infringe on fundamental rights. Algorithms exploit numerous factors that might appear unpalatable considering the significant negative consequences to individuals from risk-informed decision-making.

1. *Sociodemographic Variables*

Tools often incorporate multiple factors—directly or by proxy—implicating sociodemographic status (e.g., criminal history, education, employment, housing, family stability). By one observer's estimate, 10% to 25% of factors in risk tools are sociodemographic in nature.[313] This estimate may be low. The PSA from the Arnold Foundation, for example, is entirely populated with sociodemographic characteristics (i.e., age and criminal history).

The use of at least certain sociodemographic variables raises equal protection questions. The Equal Protection Clause embodies the philosophy that persons who are similarly situated ought to be treated alike.[314] The right exemplifies the central concepts that individuals should be accorded fair treatment in the exercise of fundamental rights and that distinctions between groups based on impermissible criteria should be prohibited.[315]

An article in the *Stanford Law Review* lays out arguments for why the inclusion of sociodemographic characteristics may be unconstitutional.[316] The author of the article disparages the vision of evidence-based sentencing practices as hardly progressive, contending current methods of risk assessment are unconstitutional when they incorporate variables implicating race, gender, or socioeconomic status.[317] As for socioeconomic-related considerations, she maintains that such factors as employment, education, income, and reliance on governmental assistance are constitutionally suspect, with arguments interweaving equal protection and due process law.[318]

Few cases challenging risk assessments on constitutional grounds exist to date. One outlier is the *Loomis* case from Wisconsin. There, the Wisconsin Supreme Court upheld the use of gender-specific algorithms in the COMPAS tool. It ruled that arrest data consistently showing women to have lower recidivism rates meant that a gender-neutral system would otherwise unfairly prejudice females.[319] Several other instruments also incorporate gender, though they vary in how they do so. The Static Risk Assessment includes gender as a predictor, which operates to reduce the score for women.[320] Similarly, the Virginia Nonviolent Risk Assessment required in sentencing

[313] Gwen van Eijk, *Socioeconomic Marginality in Sentencing*, 19 PUNISHMENT & SOC'Y 463, 466 (2017).

[314] City of Cleburne v. Cleburne Living Ctr., 473 U.S. 432, 439 (1985).

[315] Melissa Hamilton, *Risk and Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 242 (2015).

[316] Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014).

[317] *Id.*

[318] *Id*. at 830–36.

[319] State v. Loomis, 881 N.W.2d 749, 767 (Wisc. 2016).

[320] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 573-74 (Faye S. Taxman ed., 2017).

uses gender in which men are given a higher score.[321] The Ohio Risk Assessment System uses higher cut-points for women, such that at the same scores, men may be rated medium risk while women low risk.[322] Similar to COMPAS, MnSTARR and Strong-R use gender-specific calibrations for scoring.[323] The U.S. Department of Justice's new risk assessment tool—mandated by the First Step Act to be used in the federal Bureau of Prisons—is expected to have some risk factors unique to women and a separate scoring sheet by gender. Another option for developers is to use gender-specific instruments, such as the Female Additional Manual to the HCR-20, the Women's Risk Needs Assessment,[324] or the Gender Informed Supplement to the LS/CMI.[325]

To date, just one fundamental rights argument exists in the available appellate case law. In a case styled *People v. Osman*, the defendant argued that the Static-99 tool improperly assigned points for never having lived with an intimate partner for at least two years.[326] Osman claimed that the tool violated his First Amendment right regarding freedom of religion because he was single and his faith as a devout follower of Islam prohibited him from living with a lover prior to marriage.[327] Rejecting this challenge, the court upheld the actuarial scoring because the state maintained a secular purpose of identifying likely recidivists and the tool did not expressly appraise religious faith.[328]

Case-law evidence of counsel challenging risk tools for the inclusion of factors implicating sociodemographic characteristics is in its infancy. It appears reasonable to expect this area of the law to develop in the near future as algorithmic risk assessment proliferates.

As noted later herein, the incorporation of at least some sociodemographic traits may violate state law or policies, depending on the legal context of the decision, such as bail or sentencing.[329]

There is evidence that at least some of the scientists responsible for algorithmic tools are cognizant of these sociodemographic issues and have appeared to succumb to internal or

[321] Esther FJC van Ginneken, *The Use of Risk Assessment in Sentencing, in* PREDICTIVE SENTENCING: NORMATIVE AND EMPIRICAL PERSPECTIVES 9 (Jan W de Keijser et al., eds. 2019).

[322] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 573-74 (Faye S. Taxman ed., 2017).

[323] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 573-74 (Faye S. Taxman ed., 2017).

[324] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 573-75 (Faye S. Taxman ed., 2017).

[325] Emily Salisbury et al., *Gender-Responsive Risk and Need Assessment*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 355, 368 (Faye S. Taxman ed., 2017).

[326] People v. Osman, No. H037818, 2013 Cal. App. Unpub. LEXIS 2487, at *4 (Cal. Ct. App. Apr. 8, 2013).

[327] *Id.*

[328] *Id.* at *10–12.

[329] *Infra* text at note 375.

external pressures. The developers of an actuarial risk tool for sentencing purposes noted they intentionally excluded race and ethnicity as variables, vaguely referring to "stakeholder sensitivities."[330] The developers of HCR-20 were forthright about the matter: "Some risk factors, despite showing statistical associations with violence in the population, may be considered *prima facie* objectionable to include in an assessment for the purpose of estimating violence risk. Examples include race, gender, and minority ethnic status."[331] Virginia officials developed the state's own risk instrument, in the end intentionally excluding race as a rated variable, despite its statistically significant correlation with recidivism; interestingly, their justification was based on race as a proxy for social and economic disadvantage rather than the reverse.[332] As another example of political concern, the creators of the federal system's Post Conviction Risk Assessment (PCRA) purposely removed gender from the final instrument, even though their original regression model found that being female was statistically significant as a negative predictor of recidivism.[333]

Developers who are so motivated generally have reacted by resorting to regrettably unsophisticated and unempirical methods by merely eliminating ethically questionable predictors without clearly compensating for their lost predictive value.[334]

2. *Immutable Characteristics*

Alternative arguments draw on equitable principles. Critics contend that it is unethical to use risk tool results in criminal justice decisions that rely on immutable characteristics for which the individual has no control (e.g., race, ethnicity, religion, gender, and perhaps age).[335] Another critic believes the idea of "[p]aying a penalty justified only by an immutable personal characteristic runs counter to nationwide trends in equity and imposes serious societal costs."[336] Such costs include detaching punishment from the culpable act, segregating individuals within predictive groups, and suffering many false positives.[337] Other experts extend the argument to characteristics that individuals may have little control over, such as mental or physical health status.[338]

[330] Richard Berk & Justin Bleich, *Forecasts of Violence to Inform Sentencing Decisions*, 30 J. QUANTITATIVE CRIMINOLOGY 79, 87 (2014).

[331] Kevin S. Douglas et al., *Historical-Clinical-Risk Management-20, Version 3 (HCR-20^{V3}): Development and Overview*, 13 INT'L J. FORENSIC MENTAL HEALTH 93, 96 (2014).

[332] Richard P. Kern & Meredith Farrar-Owens, *Sentencing Guidelines with Integrated Offender Risk Assessment*, 16 FED. SENT'G REP. 165, 165 (2004).

[333] James L. Johnson et al., *The Construction and Validation of the Federal Post Conviction Risk Assessment (PCRA)*, 75(2) FED. PROB. 16, 19 tbl.1 (2011). PCRA creators simply noted that subsequent analyses determined the variable involving gender did not sufficiently improve the predictive validity of the model overall. *Id.* at 22.

[334] Stephen D. Gottfredson & Laura J. Moriarty, *Statistical Risk Assessment: Old Problems and New Applications*, 52 CRIME & DELINQ. 178, 194 (2006).

[335] Brian Netter, *Using Groups Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program*, 97 J. CRIM. L. & CRIMINOLOGY 699, 716-17 (2007).

[336] Brian Netter, *Using Groups Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program*, 97 J. CRIM. L. & CRIMINOLOGY 699, 728 (2007).

[337] Brian Netter, *Using Groups Statistics to Sentence Individual Criminals: An Ethical and Statistical Critique of the Virginia Risk Assessment Program*, 97 J. CRIM. L. & CRIMINOLOGY 699, 728 (2007).

[338] CHRISTOPHER SLOBOGIN, PROVING THE UNPROVABLE: THE ROLE OF LAW, SCIENCE, AND SPECULATION IN ADJUDICATING CULPABILITY AND DANGEROUSNESS 113 (2007); Thomas Nilsson et al., *The Precarious Practice of Forensic Psychiatric Risk Assessments*, 32 INT'L J. L. & PSYCHIATRY 400, 406 (2009).

In an alternative framing of such an argument, it may seem unfair to infringe on one's liberty interests based on characteristics for which the individual bears no responsibility.[339] Stakeholders perceive additional inequities whereby minorities may score higher on certain predictive factors based on reasons largely beyond their control, such as environmental phenomena that inhibit educational attainment or gainful employment.[340] The Supreme Court has given some rhetorical support for these arguments. In an equal protection opinion in which it found that classifications are not *per se* invalid by dividing classes on the basis of an immutable characteristic, the Supreme Court lamented that such divisions are contrary to our deep belief that "legal burdens should bear some relationship to individual responsibility or wrongdoing."[341]

3. Dehumanization

A related argument concerns the dehumanizing nature of algorithmic justice. Algorithmic risk tools assess dangerousness based on their selected predictive characteristics. This practice seems to objectify individuals for their traits or circumstances, which in turn undermines human dignity.[342] Instead of treating individuals as "moral *subjects*," algorithms treat them as "*objects* to be sifted, sorted, scored, [and] herded."[343] Notably, algorithmic tools are expressly used in order to so sift, sort, and score. Criminal justice algorithms are also meant to herd human beings in terms of identifying those to be arrested, incarcerated, designated to specific institutional placements, and subject to infantilizing supervisory conditions.

A legal scholar conceptualizes the role of respecting human dignity using procedural due process terms and focusing on whether a defendant perceives that the criminal justice system is treating him fairly.[344] Four aspects consider whether the process (a) treats individuals with respect and dignity, (b) presents as neutral, (c) is trustworthy, and (d) allows the individuals to substantively participate in it.[345] Algorithm-driven outcomes may violate those principles, respectively, when (a) individuals are treated like numbers rather than humans, (b) the tools are known to rely on preexisting biases and prejudices, (c) defendants cannot understand the reasoning for the algorithm's output, and (d) the process precludes defendants from making their own cases to influence decisions made about themselves.[346] Proposed solutions to improve defendants' belief in the fairness of algorithmic processes are to increase transparency of the

[339] Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 FED. SENT'G REP. 167, 171 (2014).

[340] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 430 (2016).

[341] Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 360–61 (1978) (citation omitted) (quoting Weber v. Aetna Casualty & Surety Co., 406 U.S. 164, 175 (1972)).

[342] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 19 (2019).

[343] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, EUROPEAN COMMISSION, ETHICS GUIDELINES FOR TRUSTWORTHY AI 10 (April 18, 2019), https://ec.europa.eu/futurium/en/ai-alliance-consultation.

[344] Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 UC DAVIS L. REV. 1067, 1077 (2018).

[345] Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 UC DAVIS L. REV. 1067, 1080 (2018).

[346] Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 UC DAVIS L. REV. 1067, 1080-86 (2018).

algorithms and require some semblance of human interaction between the human authority and the defendant.[347]

Dehumanization is quite perverse to long-standing sentencing traditions. For much of the last few centuries in Western cultures, the act of sentencing offenders has emphasized individualization, with the defendant being treated fairly and with dignity.[348] The ability of the decision maker to fashion an individualized sentence has been envisioned as progressive, rather than necessarily representing heedless inconsistency or lawlessness.[349] The offender is provided an opportunity to offer a personal history, to contextualize the underlying crime, and to be envisioned as having a personal future.[350] The use of actuarial assessments operates as an affront to such a system whereby the statistical tools exchange narrative with numbers, with defendants being objectified and forced into neatly contrived risk bins.[351]

The vision of removing the human from the equation operates, as well, with respect to criminal justice officials responsible for making decisions. "Heavy reliance on automated systems can alter people's relationship to a task by creating a 'moral buffer' between their decisions and the impacts of those decisions."[352] Automation allows decision makers to relinquish a sense of responsibility and accountability for their decisions by ceding their authority to the algorithm.[353] But these consequences reduce the decision makers' own humanity in the process.

B. Due Process and Equitable Considerations

This section reviews emerging topics in the law, policy, and ethics of algorithmic risk assessment practices and the challenges that lie ahead.

1. The Adversarial Process

Significant issues remain concerning the scope of an individual's right to have information regarding one's own risk assessment and to challenge various scientific and legal aspects of the prediction and, more broadly, the tool itself. The answers likely depend on the specific decision involved, relevant constitutional considerations, and the nature and degree of the consequences to the individual. Constitutional protections and case-law analysis provide for varying levels of procedural and substantive due process rights across decision points, and courts generally consider a pretrial defendant to be owed greater due process protections than a post-conviction prisoner. A sentencing decision enjoys greater due process coverage than an institutional

[347] Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 UC DAVIS L. REV. 1067, 1087 (2018).

[348] Rasmus H. Wandall, *Actuarial Risk Assessment. The Loss of Recognition of the Individual Offender*, 5 L. PROBABILITY & RISK 175, 179 (2006).

[349] Rasmus H. Wandall, *Actuarial Risk Assessment. The Loss of Recognition of the Individual Offender*, 5 L. PROBABILITY & RISK 175, 179 (2006).

[350] Rasmus H. Wandall, *Actuarial Risk Assessment. The Loss of Recognition of the Individual Offender*, 5 L. PROBABILITY & RISK 175, 189 (2006).

[351] Rasmus H. Wandall, *Actuarial Risk Assessment. The Loss of Recognition of the Individual Offender*, 5 L. PROBABILITY & RISK 175, 189 (2006).

[352] Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, presentation at Conference on Fairness, Accountability, and Transparency (2019), https://scholar.harvard.edu/files/19-fat.pdf.

[353] Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, presentation at Conference on Fairness, Accountability, and Transparency (2019).

placement classification. Civil commitment enjoys more process safeguards compared with a probationary program assignment. The comparisons could go on.

A concern is that with the algorithm-generated score presenting as scientific and factual, "predictive technology becomes another witness against the defendant without a concomitant opportunity to test the data, assumptions, and even prejudices that underlie the conclusion."[354] As a result, an expert in risk assessment practices in pretrial bail hearings reports that informed defense lawyers are negotiating to get access to any information they can on risk assessment outcomes for their clients.[355]

Case-law development generally regarding the right to receive information and to challenge assessments is a bit more advanced in sentencing. Defendants may not have a right of access to *all* the information on which the sentencing decision maker based its decision,[356] but due process requires that information relied on in sentencing be relevant, reliable, and accurate.[357] A sentence formed on materially untrue assumptions about the defendant's criminal history violates due process.[358] Thus, defendants must be given the factual information on which their sentencers relied and a meaningful opportunity to rebut it.[359] It is also noted that a sentence must normally be vacated where the sentencing judge relies on prejudicial pre-sentence material from unidentified sources that the defendant was not given an opportunity to rebut.[360] A black-box algorithmic tool might well qualify as an unidentified source with little rebuttal opportunity.

A federal case opinion recognizes that information on which a judge relies in determining the defendant's potential for future dangerousness ought to be disclosed.[361] Even staunch advocates of the use of risk assessments at sentencing promote due process protections to ameliorate potential injustices. One commentator suggests that the value of risk tools in improving on human instinct is important, but only as long as the tools are sufficiently reliable and defendants are given the ability to challenge them.[362] The Indiana Supreme Court has pontificated a bit on a defendant's access to a risk assessment's scoring sheet completed by probation as part of its pre-sentence investigation:

> A defendant is entitled to a copy of the pre-sentence report prior to his sentence being imposed . . . . Thus the defendant will be aware of any test results reported therein and may seek to diminish the weight to be given such test results by presenting

[354] Judge Noel L. Hillman, *The Use of Artificial Intelligence in Gauging the Risk of Recidivism*, JUDGES J. (Jan. 1, 2019), https://www.americanbar.org/groups/judicial/publications/judges_journal/2019/winter/the-use-artificial-intelligence-gauging-risk-recidivism.

[355] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[356] Stewart v. Erwin, 503 F.3d 488, 495 (6th Cir. 2007); United States v. Curran, 926 F.2d 59, 62 (1st Cir. 1991).

[357] Virgin Islands v. Yarwood, 45 V.I. 68, 77 (V.I. 2002).

[358] Townsend v. Burke, 334 U.S. 736, 740–41 (1948).

[359] United States v. Millán-Isaac, 749 F.3d 57, 70 (1st Cir. 2014); Smith v. Woods, 505 F. App'x 560, 568 (6th Cir. 2012); United States v. Hayes, 171 F.3d 389, 394 (6th Cir. 1999).

[360] United States v. Huff, 512 F.2d 66, 71 (5th Cir. 1975).

[361] United States v. Hamad, 495 F.3d 241, 246 (6th Cir. 2007).

[362] Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1106 (2013).

> contrary evidence or by challenging the administration or usefulness of the assessment in a particular case.[363]

Some observers see defense counsel as playing a significant role in checking the accuracy and appropriateness of risk instrument tools.[364] One advocate provided these comments:

> You should indeed have the ability at sentencing to know the results and the assumptions of the assessment and to subpoena and cross-examine any expert who administered or created the instrument if there is any dispute. Most commonly, you may provide quality assurance on such issues as whether the variables assumed by the assessment actually apply to the offender—just as when you challenge a criminal history or other aspects of a pre-sentence investigation.[365]

Similarly, another writer, acknowledging the potential for unfairness in risk assessment–led decisions in criminal justice, has suggested that the adversarial process will encourage defense attorneys to challenge the tools, which in turn may yield improvements to the science underlying them.[366] Nonetheless, counsel is unlikely in many cases to be sufficiently prepared on their own to confront the numerous and complex scientific issues involved, particularly in the face of a lack of transparency by tool owners and government agencies. Thus, counsel may need to offer expert testimony to adequately represent their client's interests.[367]

A valid argument exists that judges should themselves play a stronger role in scrutinizing the science of risk assessment. Though there is equally a legitimate concern that forensic practitioners are failing to provide judges with sufficient information to evaluate the offered evidence.

> [The law requires] that evidence be cogent. This requires that the limitations of such assessments be iterated and subjected to judicial scrutiny. Alarmingly, demonstrable limitations of risk assessments and the instruments or techniques on which they are based are all too often simply ignored by forensic practitioners of various persuasions, if they are comprehended in the first place. And so the courts are denied the very information they should be provided with when considering the prognostications of these practitioners. This lacuna must be remedied to prevent errors in the investigatory processes being relied on and hence perpetuated in the adjudicative phase with the result that miscarriages of justice are all but guaranteed to occur.[368]

Further, the science itself is so complex and uncertain that when judges engage challenges to risk assessment tools, judges appear to focus, if at all, more on the constitutional issues than evidentiary ones.[369]

[363] Malenchik v. State, 928 N.E.2d 564, 575 (Ind. 2010) (citation omitted).

[364] Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1109 (2013).

[365] Michael H. Marcus, *Conversations on Evidence Based Sentencing*, 1 CHAPMAN J. CRIM. JUST. 61, 105 (2009).

[366] David E. Patton, *Guns, Crime Control, and a Systemic Approach to Federal Sentencing*, 32 CARDOZO L. REV. 1427, 1456-58 (2011).

[367] M. Neil Browne & Ronda R. Harrison-Spoerl, *Putting Expert Testimony in its Epistemological Place: What Predictions of Dangerousness in Court can Teach Us*, 91 MARQ. L. REV. 1119, 1211 (2008).

[368] Ian R. Coyle, *The Cogency of Risk Assessments*, 18 PSYCHIATRY PSYCHOL. & L. 270, 271-72 (2011).

[369] Thomas Nadelhoffer et al., *Neuroprediction, Violence, and the Law*, 5 NEUROETHICS 67 (2012).

ℹ *Policy Considerations:*

A structure akin to mitigation services may be appropriate in terms of defense counsel offering evidence of protective and promotive factors relevant to the individual offender that may mitigate the risk score.

2. Fit for Purpose

Stakeholders should ensure the tool is fit for purpose considering its stated goal(s) and the decision(s) it informs. An issue regarding the introduction of evidence in a legal proceeding is one of relevance, also known as fitness. The proffered evidence should assist the trier of fact in understanding a fact at issue in the case.[370] Professor Christopher Slobogin is a legal expert who questions whether algorithmic risk tools are always fit for purpose in terms of the particular legal question(s) they inform.[371] He provides several examples. First, tools that predict reoffending far into the future (e.g., two, five, or 10 years) are unlikely to be helpful when the legal decision is concerned with more immediate risk (e.g., application of sentencing guideline ranges with shorter terms). Slobogin's point here highlights the portability problem in terms of using tools across decision points for which they were not designed or calibrated. As an example, a tool that was developed to assess sexual recidivism rates of sex offenders using a follow-up period of 15 years (which Static-99 does) may be suitable when the decision is the indefinite civil commitment of sexual predators. Yet, this same tool arguably does not provide much relevant information when the issue at hand is whether to release an arrested sex offender in a pretrial context. Tools that yield even a two-year recidivist projection may not be a fit in the field for police in decisions to arrest. Presumably, only a risk of imminent offending should inform arrest outcomes.

Second, algorithmic tools do not clearly inform on less restrictive means to achieve the state's goal of crime prevention (e.g., whether instead of incarceration, a more appropriate measure may be a halfway house, GPS monitoring, or treatment). Still, to the extent a jurisdiction adopts a tool in its sentencing scheme, then the decision framework should specify how the tool's results can inform on less restrictive means. It could be that the needs assessment aspect here is more exalted than the risk prediction.

Third, risk scales may not appropriately respond to the legal standard of proof. If the legal rule requires a preponderance of evidence standard, then even a "high-risk" grouping may not be sufficient unless the base rate of reoffending in that high-risk group is at least 51%. Arguably, the percentile ought to be higher considering the 95% confidence interval likely dips far below the 50% mark. Further, it may be (improperly) presumed that when a decision requires a factual showing of dangerousness, an assessment of "high risk" by definition qualifies. Such a presumption would be particularly misguided with there being no common understanding of what high risk signifies.

A related point is whether risk tools can act as ultimate issue testimony. Critics respond in the negative, as the ultimate issue in the case (e.g., whether to incarcerate, whether to parole) is a

[370] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 591 (1993).

[371] *See generally* Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583 (2018).

value judgment for the factfinder; this sort of legal conclusion is beyond the province of forensic analysts, who should be dealing in science, not law.[372] If the decision maker presumes that a label of "high risk" equates to meeting the burden of a "more likely than not" standard, the risk tool unfortunately appropriates the ultimate issue.[373] As another observer recognizes with regard to risk assessment: "Although it uses probabilistic analysis and quantification, it is not an exact science. Indeed, all science is value-laden, and risk assessment is not different in that regard. A risk appraisal can inform but cannot answer the ultimate question."[374]

Slobogin's fourth point on fitness is that if the context is sentencing, arguably only tools that predict *serious* violence ought to inform on whether the punishment includes incarceration. Tools thereby would be poor fits for the purpose of sentencing if they count nonviolent offenses, minor violence, failures to appear, or technical violations. Indeed, a similar argument can be made for other significant decisions. Presumably the risk of only serious offending should be meaningful to judgments such as pretrial incarceration, probation or parole release/revocation, sex offender registration, and sex offender civil commitment. Unfortunately, the majority of tools are not limited to predicting serious offending, as shown earlier. Stakeholders should consider this issue in their choice of tool considering the importance of the decision the tool is meant to inform and the significance of the consequences to defendants.

Additionally, it is possible that a risk tool contains items that are contrary to state statute. Sentencing law is relevant here. Several statutes prohibit consideration of race, gender, social status, or economic status in penalty decisions.[375] Sentencing commission guidance may likewise preclude the use of such factors. To the extent the risk tool directly or by proxy incorporates such data points, the law or guidance may preclude its use. Tools incorporating such factors would thereby seem not fit for purpose as contrary to statute or guidance. The same may be true with regard to statutory limits in pretrial release decisions on which individual circumstances may or may not be considered.

As a practical matter, these tools were generally not designed to answer a variety of questions for which they are being used in the real world. As mentioned earlier, these tools cannot answer questions on whether an individual should be arrested, etc. A Wisconsin Supreme Court sentencing case in 2016 acknowledged that the specific tool was not designed to address all of the various goals of punishment.[376] The likelihood of future crime was a "poor fit" for such punishment philosophies of retributivism (with its backward focus on culpability) or general deterrence (forward-looking in deterring others).[377] Hence, the court indicated the risk score may

[372] Daniel A. Kraus & Nicholas Scurich, *Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities Between Research and Practice*, 31 BEHAV. SCI. & L. 215, 224-25 (2013).

[373] Daniel A. Krauss & Nicholas Scurich, *Risk Assessment in the Law: Legal Admissibility, Scientific Validity, and Some Disparities Between Research and Practice*, 31 BEHAV. SCI. & L. 215, 226 (2013).

[374] Erica Beecher-Monas, *The Epistemology of Prediction: Future Dangerousness Testimony and Intellectual Due Process*, 60 WASH. & LEE L. REV. 353, 411 (2003) (internal citations omitted).

[375] *E.g.*, ARK. CODE ANN. § 16-90-801(b)(3) (2018) (race, gender, social, and economic status); OHIO REV. CODE ANN. § 2929.11(c) (2018) (race, ethnic background, gender, and religion); FLA. STAT. § 921.002(1)(a) (2019) (race, gender, and social and economic status); TENN. CODE ANN. § 40-35-102(4) (2019) (race, gender, creed, religion, national origin, and social status).

[376] State v. Loomis, 881 N.W.2d 749, 7 (Wisc. 2016).

[377] 881 N.W. 2d at 769.

inform the sentencing judge generally but should not be determinative of the length or severity of the sentence.[378]

3. Multidimensional View of Risk

Importantly, one should be aware of the limited nature of what risk tools predict. This issue could be considered one of fitness as well but deserves its own emphasis. Even promoters of evidence-based practices acknowledge that a key question is measuring "the risk of what?"[379] Algorithmic tools generally predict single-event probabilities.[380] Presumably, though, the idea of risk for purposes of penalties and management decisions is not some unitary characteristic focused solely on an abstract likelihood of committing one antisocial act sometime in the future. Instead, at least six different dimensions of risk are conceivably pertinent. Probability is only one of them. Depending on the decision context, it may not be as important as the other five. As represented in Figure 17, the additional dimensions include offense type (e.g., terrorism, violent, property, white collar, drugs), severity of harm (which may overlap with crime type), imminence, frequency, and duration of offending.[381] Unfortunately, the currently available tools generally ignore most of those additional, yet important, dimensions represented in Figure 17.

Figure 17: Multidimensional Perspective on Risk

Indeed, additional dimensions may be appropriate along the lines of evidence of specialization versus versatility in offending. A significant reform here would be to include an orientation around studying the factors that predict desistence.

In contrast to the more relevant multidimensional perspective on risk, developers of risk assessment tools generally have operationalized failure as a simple dichotomous measure.

[378] 881 N.W. 2d at 768.

[379] Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 Duq. L. Rev. 707, 743 (2011).

[380] David J. Cooke & Christine Michie, *Violence Risk Assessment: From Prediction to Understanding — or From What? To Why?*, *in* Managing Clinical Risk: A Guide to Effective Practice 3, 18 (Caroline Logan & Lorraine Johnston eds., 2013).

[381] Michael H. Fogel, *Violence Risk Assessment Evaluation: Practices and Procedures, in* Handbook of Violence Risk Assessment and Treatment: New Approaches for Forensic Mental Health Professionals 41, 43 (Joel T. Andrade ed., 2009).

Actuarial tool developers usually have counted a person as a recidivist as soon as that individual in the developmental sample committed a qualifying act during the follow-up period of observation. Thus, actuarial tools might treat identically these two hypothetical people: (1) the subject who immediately upon release began a long crime spree involving heinous violent offenses causing significant harm to a variety of victims; and (2) a subject who once violated a nonserious supervisory condition out of negligence years after release. But when risk is the basis of a significant loss of freedom, it should matter not only the probability of some future act but also the magnitude of the potential harm.[382] Clearly, the actual danger, harm, and culpability posed by these two hypothetical offenders are quantitatively and qualitatively disparate, yet the actuarial tools usually fail to differentiate between them.

As another example of poor fit, tools designed to predict pretrial failure often conflate the act of intentionally absconding (e.g., fleeing the jurisdiction) with a failure to appear for involuntary reasons that do not indicate an intention to avoid trial (e.g., health emergency, failure of transportation, forgetfulness).[383] The system's goal in this respect of ensuring the defendant's appearance in court is not thwarted in the same way in those scenarios, but label bias in defining the failure event might force the algorithm to count them synonymously.

The observation that actuarial tools fail to a large degree to provide information relevant to legal decisions is not surprising considering many of the tools are exogenous to the law, having been developed in the fields of mental health and behavioral sciences.[384] The purpose for creating some of the violence risk assessment tools, for example, was to prevent recidivism through clinical interventions; they were not originally intended to allow for predictions of dangerousness that may trigger legal consequences.[385] To highlight: VRAG, a popular violence risk tool, was created to determine the treatment needs of violent offenders with mental disorders and to inform decisions about which psychiatric patients in a secure hospital facility should be released.[386] Tools designed for one purpose may be inappropriately morphed into criminal justice; this is concerning considering the complex and multifaceted justice decisions that require weighing various legal and extralegal factors.[387]

4. Potential Overreliance on Criminal History

The tools disproportionately rely on criminal history. Rarely do they use anything analogous to a statute of limitations. In other words, the tools tend to count criminal history no matter how ancient.[388] Colloquially, most tools do not "clean the slate,"[389] even of juvenile offenses or

[382] Christopher Slobogin, *The Modern Case for Indeterminate Dispositions in Criminal Cases*, 48 SAN DIEGO L. REV. 1127, 1135 (2011).

[383] PARTNERSHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE System 17 (2019).

[384] BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 188 (2007).

[385] Michael H. Fogel, *Violence Risk Assessment Evaluation: Practices and Procedures, in* HANDBOOK OF VIOLENCE RISK ASSESSMENT AND TREATMENT: NEW APPROACHES FOR FORENSIC MENTAL HEALTH PROFESSIONALS 41, 42 (Joel T. Andrade ed., 2009).

[386] VERNON L. QUINSEY ET AL., VIOLENT OFFENDERS: APPRAISING AND MANAGING RISK 25 (1998).

[387] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 279 (2013).

[388] *See generally* Melissa Hamilton, *Back to the Future: The Influence of Criminal History of Risk Assessments*, 20 BERKELEY J. CRIM. L. 76 (2015).

[389] Faye S. Taxman, *Risk Assessment: Where Do We Go From Here*?, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 271, 275 (Jay P. Singh et al. eds., 2018).

acquitted conduct. This stance is concerning as the behavioral science evidence is to the contrary in terms of past offending having some limit to predictive ability.

> [T]here is a redemption point, or a time when an individual presents the same risk for arrest (proxy for criminal behavior) as the general population at the same age; that is, age becomes a function of risk for reoffending. To summarize briefly, depending on the age of first arrest, there is a different point of redemption for different types of offenses.[390]

Even sex offenders, though widely presumed to be lifelong recidivists, can desist. Experts point to evidence that sex offender recidivism is highest in the first few years then decreases substantially thereafter.[391]

Care must be taken, too, about the same criminal history event(s) being counted numerous times.[392] For some of the tools, one charge may be counted within multiple risk factors. This duplicative counting may be exacerbated further if the decision maker is already considering prior offending. For example, criminal history is a significant component of sentencing decisions. If the risk tool used also heavily relies on measures of past offenses, the role of criminal history may become overly exaggerated, leading to disproportionate responses or penalties.

It is possible that the incorporation of criminal history into tool factors might violate a relevant statute or policy. For example, if a sentencing guideline system excludes certain evidence from its criminal history calculations (e.g., juvenile offenses, arrests without convictions, or foreign convictions), then, arguably, this type of evidence should not pervade the sentencing decision through the back door of a risk tool.

A similar warning may apply to other information types included in any particular tool. Double counting may, for example, be occurring with respect to the offender's unemployment status, housing instability, and antisocial attitudes. Still, the likelihood of repeated use by the tool and then separately by the decision maker of these other considerations is not as salient as criminal history because of the latter's unequaled presence and weight in most criminal justice risk tools.

5. Need for Expert Evidence

Algorithmic risk assessment is a convoluted, technologically savvy, and ever-changing regime. Few people understand much about these algorithms. Relevant information needed includes the following:

- methodologies for how the tool and algorithm were created
- details on the algorithm, its factors, and weights
- the codebook and any scoring sheets

[390] Faye S. Taxman, *Risk Assessment: Where Do We Go From Here*?, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 271, 275 (Jay P. Singh et al. eds., 2018).

[391] *See generally* R. Karl Hanson et al., *High-Risk Sex Offenders may not be High Risk Forever*, 29 J. INTERPERSONAL VIOL. 2792 (2014).

[392] *See generally* Melissa Hamilton, *Back to the Future: The Influence of Criminal History of Risk Assessments*, 20 BERKELEY J. CRIM. L. 76 (2015).

- definition of criminal justice failure used (e.g., any arrest, supervision failure)
- details on the sociodemographic makeup of the training, developmental, and validation samples
- any evidence of validation, the criteria used, and results
- training materials
- the sites and sociodemographic makeup of population(s) of cross-validations
- information on how the tool operates in practice
- embedded value judgments (e.g., decisions on grouping scores; thresholds for low, moderate, high risk)
- identification of the factors therein that may act as proxies to sociodemographic characteristics (e.g., correlation coefficients)
- whether and how the tool is biased using the various group fairness definitions available
- information on attempts to reduce biases
- inter-rater reliability scores for evaluators
- information on how much human involvement is retained
- information on override policies, whether policy or professional (including any oversight mechanisms for overrides)
- number and nature of overrides and rates of them applying to sociodemographic characteristics and offense types

Confusion reigns for all tools, proprietary or not. Even governmental agencies who have created their own risk tools are notoriously secretive about sharing much information about them or their data.

As a result, defense attorneys likely need the assistance of expert witnesses with specialized knowledge to understand the risk tools and to properly confront potential scientific flaws and ethical issues. Subject matter experts can explore issues in design, accuracy, proper validation, implementation, and qualifications of developers and evaluators. Expert evidence is required about the various aspects of the tool itself and about whether the particular client was properly scored. Accurate scoring may be an issue when, while some input factors may be relatively objective, others require subjective evaluation (e.g., presence of a mental disorder, elementary school maladjustment, criminal thinking style, having criminal friends). Forensic expertise may be helpful for deciphering codebooks and byzantine scoring equations.

Still, arguments may be fruitful regarding even the more “factual” aspects. It could be the criminal history record used to score the tool is erroneous (e.g., “that was my father”; “I was acquitted”). The evaluator may have committed a data-processing error (e.g., input the wrong birthdate) or incorrectly summed points on a hand-scored instrument. Or an explanation may justify rescoring the tool (e.g., “the offense was 50 years ago, and I was a minor”; “I did not show up for my court date but there was an unavoidable emergency that reasonably explains the failure to appear”). An expert may be useful in determining whether the individual’s score constituted an override and to challenge its legitimacy on that front.

Notwithstanding the need for expert assistance, this is another way the justice system may advantage wealthier defendants, who are more likely to be able to afford such expertise. This raises civil rights issues regarding differential access to justice for the rich versus the poor. Still, to the extent challenges are successful, poor defendants may eventually benefit therefrom.

ℹ *Policy Considerations:*

Counsel should make use of any form of discovery that is available at the particular decision point to gain as much information as possible about the tool itself and the defendant's scoring.

Motions for expert witnesses at the expense of the state may be appropriate considering the nontransparency of risk algorithm processes.

In appropriate contexts, criminal justice agencies could make available to defendants neutral subject matter experts who can explain relevant aspects of the specific algorithmic tool in terms of its development, modification, operation, validation, and biases. Funding for these experts could be included within the implementation and maintenance budget.

Teams could develop a knowledge library to share information about specific tools, risk assessment practices, ideas about what works best, and unintended consequences of otherwise well-intentioned policies.

More continuing legal education trainings could usefully be offered regularly and updated as risk assessment progresses into new, perhaps more technologically heavier regimes. Such training should include statistical and empirical research skills.

6. Self-Incrimination

Many of the tools require an interview with the offender to obtain data to score them. Generally, evaluators question offenders for these purposes outside the presence of counsel and quite often without their lawyer's knowledge. Yet queries that are needed in order to address predictive factors embedded in the tools may solicit responses that are self-incriminating. Commonly, offenders are asked to:

- identify prior undetected crimes
- admit to antisocial tendencies
- acknowledge drug/alcohol problems
- provide information indicating that one's livelihood depends on proceeds from crime (e.g., a criminal livelihood)
- reveal information about personal relationships, such as having criminal friends or criminal family members

At least one tool, the Self-Appraisal Questionnaire (SAQ), elicits criminal history information entirely from the offender's self-survey reporting.[393] The SAQ asks offenders to respond to statements such as these:[394]

[393] Wagdy Loza, *Self-Appraisal Questionnaire (SAQ): A Tool for Assessing Violent and Non-Violent Recidivism, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 165, 166 (Jay P. Singh et al. eds., 2018).

[394] Wagdy Loza, *Self-Appraisal Questionnaire (SAQ): A Tool for Assessing Violent and Non-Violent Recidivism, in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 165, 166 (Jay P. Singh et al. eds., 2018).

- "I have carefully planned a crime before."
- "My criminal involvement has been getting worse."
- "I would not have served time if it was not for my alcohol or drug habit."

The SAQ guidance suggests that an evaluator who finds any discrepancies between the individual's responses and their official record should ask the individual for an explanation.[395]

As another illustration, the federal Post Conviction Risk Assessment tool includes a survey section for offenders to complete. Examples of queries in the offender's survey from the Psychological Inventory of Criminal Thinking Styles (PICTS) include the following:[396]

- "When pressured by life's problems I have said 'to hell with it' and followed this up by using drugs or engaging in crime."
- "I have found myself blaming the victims of some of my crimes by saying things like 'they deserved what they got' or 'they should have known better.'"
- "The way I look at it, I've paid my dues and am therefore justified in taking what I want."
- "The more I got away with crime the more I thought there was no way the police or authorities would ever catch up to me."
- "I believe that breaking the law is no big deal as long as you don't physically hurt someone."
- "I have helped out friends and family with money acquired illegally."
- "Nobody tells me what to do, and if they try, I will respond with intimidation, threats, or I might even get physically aggressive."
- "There have been times when I have felt entitled to break the law in order to pay for a vacation, new car, or expensive clothing that I told myself I needed."

Both the PICTS and SAQ have built-in trick questions designed to detect deception and self-presentation biases.[397] The PICTS author describes them as signs of faking good or faking bad in responses.[398]

Policy Considerations:

Jurisdictions should ban practices in risk assessment interviews of demanding waivers of confidential information.

Requests for confidentiality waivers should involve a right to consult with counsel.

Questions that might elicit self-incriminating information should be excised.

[395] Wagdy Loza, *Self-Appraisal Questionnaire (SAQ): A Tool for Assessing Violent and Non-Violent Recidivism*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 165, 168 (Jay P. Singh et al. eds., 2018).

[396] PICTS (on file with author).

[397] Wagdy Loza, *Self-Appraisal Questionnaire (SAQ): A Tool for Assessing Violent and Non-Violent Recidivism*, *in* HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 165, 173 (Jay P. Singh et al. eds., 2018).

[398] Glenn D. Walters, *Predicting Recidivism with the Psychological Inventory of Criminal Thinking Styles and Level of Service Inventory-Revised: Screening Edition*, 35 LAW & HUM. BEHAV. 211, 217 (2011).

7. Validation Issues

One American court has directly addressed the legality of a criminal justice decision if it is based on an algorithmic tool that was not properly validated. The defendant in *Wisconsin v. Loomis* brought a due process challenge as his pre-sentence report contained information from an algorithmic risk assessment. Loomis argued that the proprietary nature of the specific tool (COMPAS) prevented him from contesting its scientific rigor.[399] The Wisconsin Supreme Court ruled the defendant had the opportunity to challenge his risk outcome to the extent he had been able to refute the information on which his score was based as the data were either public record or derived from his own interview responses.[400] The *Loomis* court was unconcerned that the defendant was not able to obtain any intelligence from the algorithm itself, upholding its trade secret protection. Nonetheless, the majority acknowledged the existence of scientific issues with algorithmic risk tools. Instead of prohibiting their use, however, this court's chosen remedy required that a written list of cautions accompany the risk results, tailored specifically to the COMPAS tool. The cautions include the following:

> * The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined.
>
> * Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.
>
> * Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.
>
> * A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations
>
> * COMPAS was not developed for use at sentencing but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.[401]

The opinion correctly acknowledged the time sensitivity of these cautions. It admonished criminal justice officials to stay current on scientific developments and to "continuously assess" the proper use of any risk assessment tool.[402] The majority opinion surmised that if a validation study is completed in Wisconsin or if other information about risk assessment practices becomes available, these cautions may become more or less relevant, and thus require updating.

Courts in other countries have ventured further on the validation issue. The Canadian Supreme Court in 2018 ruled that it was unfair to use an algorithmic tool on an Indigenous prisoner because there was no evidence that the particular tool had been validated specifically

[399] State v. Loomis, 881 N.W.2d 749, 753 (Wisc. 2016).

[400] *Id*. at 771.

[401] *Id*. at. 769-70.

[402] *Id*. at 753.

on the Indigenous Canadian population.[403] Likewise, at least one court in Australia refused to consider an algorithmic risk score for an Indigenous offender because of the lack of any relevant validation study.[404] However, other Australian courts have been more pragmatic in being willing to consider algorithmic predictions of risk for Indigenous Australians, despite recognizing the lack of validation on such groups weakened such evidence.[405]

Challenges on grounds of appropriate validations are only likely to flourish in the future as algorithmic risk assessment both expands and attracts more attention. Still, the Canadian case brings to mind a related, though untouched, issue. What equal protection or due process issues arise if an agency in their decision-making uses an algorithm on some individuals/groups within its population but not others because of validation issues? Does the group with algorithmic-assisted decision-making have a valid equal protection claim to reject the use of algorithms? Or does the group that is left with completely human decision-making have a claim? Do they potentially both have valid claims?

ℹ *Policy Considerations:*

Authorities should carefully craft written lists of limitations that are honed to the specific tool, context (e.g., pretrial bail decision, sentencing, post-release programming), and intended population.

Mandates on data retention practices for algorithmic tool development and modification are necessary to ensure defendants have access to information about the tool, its development, its training data, the algorithm, and any updates or modifications thereto necessary to allow meaningful review.

Mandates on data retention practices at decision points are necessary to permit the defendant access to the data inputs, tool outcomes, and overrides that are applicable in the individual case to allow meaningful review and contest the individual score and outcome.

Due process protections at important decision points require an evidentiary hearing and an appropriate level of discovery for the individual's assessment concerning the tool, information relied upon, and the scoring and if an override applied.

8. Improper Delegation

In multiple ways, risk assessment practices may improperly cede authority to the wrong people.[406] An agency or jurisdiction that adopts a risk assessment tool may appear to be engaging in rulemaking yet without appropriate safeguards or oversight that would normally govern official rulemaking agencies.[407] Importantly, in many jurisdictions, judges are given statutory

[403] Ewert v. Canada, 2018 S.C.R. 30, para. 66 (S.C.C. June 13, 2018).

[404] DPP v Samson (WA) [2014] WASC 199 (Austl.).

[405] Alfred Allan et al., *Assessing the Risk of Australian Indigenous Sexual Offenders Reoffending*, 26 PSYCHIATRY PSYCHOL. & L. 274, 284 (2019) (listing cases).

[406] Sarah Valentine, *Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L. J. 364, 372 (2019).

[407] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 22 (2019).

discretion to make certain decisions (e.g., bail, sentencing). Yet such judicial discretion may be infringed on by the actions of police, prosecutors, or tool developers in controlling and manipulating risk tools, their operations, outcomes, and overrides.[408] This would exemplify the executive branch interfering with the judicial branch's authority.[409] Hence, where the law places discretionary authority in the hands of one body, it might be illegal and/or unethical to delegate that discretion to another body—even a nonhuman algorithm.[410]

Anecdotally, an outcome of "high risk" from an algorithmic tool invariably leads to denial of parole or to commitment in sexual predator civil commitment proceedings.[411] In these cases, the person creating the category of "high risk" appears to be the one who has subsumed authority for those outcomes. As another example, suggest a state law mandates that a judge grant an offender pretrial release unless the judge determines the individual is at a high risk of failing to appear. The determination of a tool's cutoff for "high risk" may interfere with judicial authority in such a case. Similarly, the tool developer's choices on cost ratios (preferring false positives over false negatives (or vice versa), and the extent thereof, may constitute an authoritative power grab from the judiciary.[412] Even the basis for algorithmic tools can be envisioned as incompatible with discretionary decision-making. Discretion, by its very nature, cannot be tied to a fixed set of automated rules.[413]

9. Hypothetical Future Offense

A cornerstone of the American criminal justice system is the presumption of innocence. Critics of a risk-based criminal justice system, one in which predictions can dictate sanctions or restrictions, charge that such a system inherently results in punishing an individual for potential future behavior.[414] That is, risk predictions might constitute a criminalization of the hypothetical crime (i.e., a precrime as presented in the film *Minority Report*). And with risk technologies being developed on group-based data—and thus deindividualized—the scheme has been described by a reporter as merely representing "mechanical crime prediction."[415]

In a sentencing context, punishing for the possibility of reoffending disrupts certain punishment theories.[416] The practice denies the specific deterrence ability of the immediate conviction and sentence as well as negates the assumption of free will. Humans are fundamentally unreliable beings. Human behavior is hard to predict as humans are active, reactive, interactive, and adaptive creatures. There is absolutely no certainty whether a person

[408] Discussion at Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[409] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[410] Marion Oswald, *Algorithm-Assisted Decision-Making in the Public Sector*. 376 PHIL. TRANSACTIONS 1, 14 (2018).

[411] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 559 (2018).

[412] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 559 (2018).

[413] Marion Oswald, *Algorithm-Assisted Decision-Making in the Public Sector*. 376 PHIL. TRANSACTIONS 1, 14 (2018).

[414] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 277 (2013).

[415] Leon Neyfakh, *You Will Commit a Crime in the Future: Inside the New Science of Predicting Violence*, BOSTON GLOBE, Feb. 20, 2011.

[416] Michael Marcus, *MPC-The Root of the Problem: Just Deserts and Risk Assessment*, 61 FLA. L. REV. 751, 753 (2009).

will or will not commit a future act. Hence, tools that assign probabilities to extremely low- and high-risk groups of 0% and 100%, respectively, represent statistical fictions.

An alternative frame of this problem of aggravating punishment for a hypothetical, future offense is to conceptualize it as justifying a framework of inchoate crimes, but without requiring criminal law's otherwise fundamental elements of proving a culpable mental state (mens rea) and some corresponding conduct (actus reus). The crime is merely hypothetical in these scenarios, yet the consequences are individually experienced as real.

10. Punishing the Individual for Group Behavior

Actuarial risk practices amount to punishment based on shared group characteristics.[417] The G2i issue is relevant here. Former United States Attorney General Eric Holder warned against using algorithmic tools to inform sentencing decisions because "[u]sing group data to make an individualized determination . . . can result in fundamental unfairness."[418] Another observer argues that the "law, which aims to effect justice, is understandably resistant to determining one individual's fate on the basis of data drawn from others, no matter how large or representative the sample."[419] It is tantamount to punishing someone for the crimes that other persons—alleged to be statistically matched—have committed in the past.[420] Even this last conceptualization is overly generous. Because the actuarial tools often count arrests and other proxies to crime, it may be that many of the labeled recidivists did not offend.[421]

Recidivism rates in the training data for other reasons may also be overstated. This observation is a part of the broader issue of the significant false positive rates by using actuarial estimates, particularly as the tools tend to overcorrect to avoid false negatives. The question then is the fairness of negative consequences in a regime where many false positives inevitably occur.

ℹ *Policy Considerations:*

Communication standards should clarify the group-based nature of the risk assessment project and that the results are relative to a group (with appropriate descriptors) and not absolute to the individual.

Evaluators should provide 95% confidence intervals if they offer estimates of percentages normed on the developmental samples to make it clearer the variability of the statistics.

[417] Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 277 (2013).

[418] Joshua Barajas, *Holder: Big Data is Leading to 'Fundamental Unfairness' in Drug Sentencing*, PBS NEWS HOUR (July 31, 2014, 6:29 PM).

[419] ANDRE A. MOENSSENS ET AL., SCIENTIFIC EVIDENCE IN CIVIL AND CRIMINAL CASES 1259 (5th ed. 2007).

[420] J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1390 (2011).

[421] Christopher Slobogin, *Dangerousness and Expertise*, 133 U. PA. L. REV. 97, 123 (1984).

11. Status Offense

An alternative construction for the idea of sanctioning the hypothetical crime or for group behavior is to conceive of the issue as one of penalizing an individual for his/her status.[422] As the Supreme Court recently averred: "Our law punishes people for what they do, not who they are."[423] Here, the criminalizing status (e.g., "high risk") is one presumed to be indicative of future dangerousness. The tendency to tolerate a high false positive rate is seen as a form of risk inflation that stigmatizes offenders deemed high risk, leads to overincarceration, and reduces emphasis on rehabilitation.[424] The "high risk" label becomes a sort of master status that criminal justice officials respond to, objectify (even reify) as proof of a deviant character, and punish as a result. Thus, the label or the risk score justify a form of criminal profiling, one that appears to be supported by a veneer of science-led algorithms.

XI. Implementation of a Risk Assessment Program

Risk assessment tools are not interchangeable or universal in application. Stakeholders must take care to select (or build) a tool designed for the site's unique goals, context, population, resource abilities, and infrastructure.[425]

> *"Inappropriate selection or implementation of risk assessments carries considerable detrimental consequences for the offender and the public at large."*
>
> Williams et al. (2001)

Attention must be paid to a host of decisions and policies that must be made, both up front and throughout implementation, in order to achieve the intended goals while avoiding unnecessary negative consequences to individuals assessed. Evidence-based practices may rely (somewhat) on the scientific method in tool development, but this takes place in a legal and political environment demanding significant stakeholder involvement, if not focused direction.

A. Operational Decisions

Scientists who develop and/or operate algorithmic tools have in practice made decisions regarding their use that actually are judgment calls that instead ought to be made or directed by policy makers (e.g., elected officials, judges, administrative agencies). As risk assessment

[422] Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 Ohio St. J. Crim. L. 583, 590 (2018).

[423] Buck v. Davis, 137 S. Ct. 759, 778 (2017).

[424] Jodi L. Viljoen et al., *Do Risk Assessment Tools Help Manage and Reduce Risk of Violence and Reoffending? A Systematic Review*, 42 Law & Hum. Behav. 181, 184 (2018).

[425] Kevin M. Williams et al., *The Use of Meta-Analysis to Compare and Select Offender Risk Instruments*, 16 Int'l J. Forensic Mental Health 1, 2 (2017).

practices have progressed incrementally, the takeover of various policy decisions has occurred, albeit without proper notice. Indeed, even the development of a tool is a policy-laden exercise in which numerous normative judgments are required.[426]

1. *Multidisciplinary Implementation Panel*

Historically, tool adoption and implementation have too often been haphazard, ill-informed, and covertly accomplished by a relatively small group of actors. Times are changing, though, as algorithmic risk practices are more widely revealed and debated. The best-practices approach now is through a more coordinated effort via a multidisciplinary team that can draw on the skills, experiences, perspectives, and needs of its varied members. The adopting jurisdiction must clearly articulate its purpose(s) for adopting a risk assessment tool.[427] Then it should impanel members who are best suited to the task of achieving these goals while protecting individual rights.

The panel has multiple responsibilities in this best-practices model. A carefully crafted implementation plan can bolster success. The plan must do many things. Among them is to identify various means to include and encourage the buy-in of evaluators who will score the tools and other end users.[428] *End users* here refers to the officials who in the field will be making the decisions that risk results inform (e.g., judges at sentencing, probation officers initiating revocation, paroling authorities granting release).

Experiences to date indicate that an impediment to success can occur when criminal justice officials fail to fully trust algorithmic risk scores and thus may be ignoring them.[429] A 2011 Kentucky law mandates risk assessment as part of judges' pretrial release decisions, with defendants who are ranked as low or moderate risk earning a presumptive release without bond.[430] If the presumption had been followed, the pretrial release rate in Kentucky was expected to have increased by 37%; but many judges refused to follow the presumption, and the actual pretrial release rate increased by just 4%.[431] Virginia adopted a risk assessment tool in 2002 to isolate the lowest risk among nonviolent offenders for diversion from incarceration.[432] In practice, the rate of incarceration rose by a percentage point.[433] The reason for both scenarios is that, while some judges followed the intended directions, many rarely did so or for only a portion of the time.[434]

[426] Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 61 (2017).

[427] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 568 (2015).

[428] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 575 (2015).

[429] Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181, 217-18 (2019) (citing evidence).

[430] Public Safety and Offender Accountability Act, H.B. 463, 2011 Gen. Assemb., Reg. Sess. (Ky. 2011).

[431] MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM 5 (2018), https://www.acslaw.org/analysis/reports/roadblocktoreform.

[432] MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM 6 (2018).

[433] MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM 7 (2018).

[434] MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM 7 (2018).

Still, these experiences do not mean that a plan to use risk assessment tools to reduce the incarcerated population is destined to fail. A bail reform bill in New Jersey that relied on risk assessment successfully allowed that state to substantially reduce its jail population by 35%,[435] and during a time when crime rates dropped.[436] The New Jersey legislature appropriately had included judges in its implementation, giving them a significant role in choosing the instrument and in crafting a decision framework dictating how risk assessment would be employed to achieve the population reduction.[437]

The lesson to be learned here is the value in encouraging and providing avenues for end users to be involved with how the tool is developed, maintained, adapted, and operated. Selecting an off-the-shelf tool is related to weak acceptance by stakeholders and poor classification quality.[438] Another reason to gain buy-in is that successful implementation will require a significant cultural shift in the criminal justice agency.

> Policymakers' need to know the subsequent strategies for public safety and recidivism reduction might begin with a simple question: Do risk assessment instruments reliably predict recidivism? The short answer, according to years and volumes of research, is resoundingly: yes. But we must be mindful of what saying yes may mean. Adoption of a risk assessment tool goes hand-in-hand with fundamentally altering approaches to reentry and correctional management, supervision, services, and more broadly criminal justice practice. Ultimately, the process of implementing risk assessments within an agency should consist of more than simply adding a tool to the agency portfolio; it should result in a shift of corrections culture, practices, and policies.[439]

The multidisciplinary panel should also be intimately involved with decisions regarding the algorithmic tool adopted. "There remains significant opportunity to influence and manage the development of computer technology, to ensure that ethics and law are part of the curriculum of software developers and analysts, and to regulate as necessary."[440]

Currently available risk tools are uninformative about much of what should be a multifaceted picture of risk. Sentiments about the multidimensional nature of the task should be subject to public debate before choosing and implementing any tool. It could be a creative design team that crafts a new model to incorporate these dimensions. The multidisciplinary team as a rule is not limited to considering off-the-shelf tools. Criminal justice agencies have been developing their

[435] MEGAN T. STEVENSON & JENNIFER L. DOLEAC, THE ROADBLOCK TO REFORM 9 (2018).

[436] PRETRIAL JUST. INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA 4 (2017); Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[437] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[438] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 536 (Faye S. Taxman ed., 2017).

[439] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 22 (Faye S. Taxman ed., 2017) (quoting from Sarah L. Desmarais & Jay P. Singh, Council of State Governments, an Executive Summary Drawn From Assessing Recidivism Risk: A Review of Validation Studies (2013)).

[440] Rónán Kennedy, *Algorithms and the Rule of Law*, 17 LEG. INFO. MGMT. 170, 172 (2017).

own tools to have control over them, and this could be an improvement on currently available commercial tools.

Throughout the rest of this section, additional ways that the panel can appropriately influence implementation will be offered.

Policy Considerations:

The jurisdiction maintaining or adopting an algorithmic tool should create, adequately fund, and sufficiently staff a multidisciplinary panel to provide oversight and a forum for debate on the many issues that can allow risk assessment practices to succeed as well as ameliorate negative consequences.

The multidisciplinary implementation panel may include, depending on the context and decision, representatives of agency personnel, end users, defense counsel, academics, prosecution, police, forensic organizations, current or former prisoners, victims' groups, and community organizations.

The multidisciplinary panel should engage in efforts before, during, and after implementation toward directing how the algorithmic tool is created and operates.

The panel should consider that, at most decision points, a tool that predicts only serious offending is likely appropriate. The panel should otherwise ensure that the tool is fit for the purpose(s) of the decision it is intended to inform.

The panel might consider if developing a tool that predicts desistence or successful reentry is desirable.

The multidisciplinary panel should make decisions on the minimum validity levels that are acceptable and which validity and group fairness measures matter more than others.

The multidisciplinary panel should, through open debate, make relevant decisions on how best to deal with sociodemographic characteristics and their proxies. These decisions must be weighed against cultural sensitivities and predictive abilities.

The multidisciplinary implementation team should consider the potential for criminal history to overwhelm decisions to an unreasonable degree. Options could be to modify the algorithm to reduce reliance on criminal history measures or to build in protections within the decision framework. Limits should be placed on the use of criminal history consistent with those existing in the legal framework outside of risk assessment. Consideration should also be given to refining criminal history to include some way to factor in the age-crime curve and the progressive loss of salience of old offenses as time passes.

A pilot study before full implementation should be conducted, if feasible.

2. *Decision Frameworks*

A decision-framework approach understands that end users must be guided in what to do with risk information. Such guidance must closely adhere to the purpose(s) for which the tool was engaged. A common aim in recent years is for a risk algorithm to help produce a higher pretrial release rate. With that goal in mind, guidance could inform judges on the connection between algorithmic scores and release decisions. Indeed, in such an instance the guidance could reframe the whole project from the implementation of a *risk* tool to that of a *release* tool.[441] The following bubble summarizes a defense advocate's clever musings on the topic:[442]

> Judges already know how to incarcerate. They don't need any further information on how to do that. What they do need is a reason for why *not to* incarcerate.

A carefully thought-out framework can respond to such a task. A decision framework might need to tackle the more difficult judgments that algorithms cannot directly dictate yet are still used to inform. These involve decision points such as whether to arrest, the in/out decision in sentencing, the length of a sentence, or parole release. Officials responsible for these determinations could helpfully use guidance on translating how the risk assessment result is useful in their thought processes. How the implementation team structures this guidance will likely depend on the jurisdiction's prevailing values considering competing theories of punishment. An option would be that risk predictions operate as presumptions of some sort of decision (e.g., low risk presumes diversion).

A decision framework may need to guide end users in how to limit overlap in risk factorology. The tools consider many factors that have already been seen as relevant pieces of information for the various decisions that risk assessment is intended to educate. The overlap raises, though, the probability that the same factors will be counted twice: once within the algorithmic score and then again with the decision maker's usual thought process in contemplating relevant evidence. For instance, if the decision maker typically considers drug abuse, and the tool also scores drug abuse, the influence of that factor may become disproportionate and unreasonably prejudice the consequences to the offender.

Conversation about risk assessment commonly orients to low risk and high risk. Relatively little attention is given to the categories in between. A decision framework could provide more guidance there. In a pretrial context, should moderate risk be considered more toward a presumption of release or a reason to deny release? Is a moderate-risk outcome suggestive of probation but with strong community supervision? Does moderate risk equate to a medium-security classification?

[441] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 13:04 start time) (on file with NACDL).

[442] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 13:04 start time) (on file with NACDL).

Then in certain decision contexts, the practice of reassessment is important. Particularly to the extent that tools contain dynamic risk factors and protective or promotive factors, valuable information may be obtained by reevaluating individuals at periodic intervals. It could be that one's risk profile has substantially changed through programming, by the person's own desire to emerge as a prosocial citizen, or by aging out. Risk assessment practices can then help decision makers recognize and reward individuals for their rehabilitative successes. Case management should be flexible to changing plans to meet the individual's altered risk profile.

Care must be taken, though, not to oversell algorithmic risk assessment.[443] "Unless criminal justice system actors are made fully aware of the limits of the tools they are being asked to implement, they are likely to misuse them."[444]

A particular fear for misuse may be of a net-widening effect. The inclusion of the needs model may appear to benefit all defendants, at least to the extent that individualized programming is assigned to improve chances of success. But unintended consequences are possible. For example, the use of risk-needs tools as a pretrial alternative to money bail has on occasion had the downside of turning pretrial release agencies into pretrial service agencies.[445] Instead of a primary focus on whether the defendant will appear for court dates, the agencies assign services, such as drug testing, drug treatment, and monitoring school or work attendance. But these sorts of mandates also increase the chances of noncompliance and thus a potential return to incarceration, which infringes on liberty and privacy interests. The decision framework could preemptively avert foreseeable abuses while rewarding compliance.

ℹ *Policy Considerations:*

The multidisciplinary panel should create a written decision framework that contains clear guidance on how the relevant decision maker/agency should use the specific tool and for what purposes.

A risk assessment tool's outcome should never autonomously dictate a result that has negative consequence to those assessed. Instead, a tool should inform but not entirely replace a human decision maker.

The decision framework should be clear that risk assessment results can inform but should not be used on their own to settle the ultimate issue.

Depending on the complexity of the decision framework, training of evaluators and end users on the framework may be appropriate.

3. Thresholds

The ranking by risk scheme often is attuned to creating or meeting certain thresholds. This could be determining which failure rates justify a low-, medium-, or high-risk ranking (e.g., 10%, 30%, 50% recidivism rates?). It might entail setting a threshold as a basis for a given action or

[443] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 576 (2015).

[444] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 576 (2015).

[445] Malcom M. Feeley, *How to Think About Criminal Court Reform*, 98 B.U. L. REV. 673, 684 (2018).

nonaction (e.g., low risk presumption of pretrial release, high risk suggests a sentence involving incarceration). A somewhat similar baseline option is the acceptability level of the false positive rate or false negative rate (or, more appropriately, the prospective measures of PPV/NPV).

With the lack of industry or legal standards for thresholds, it is often left up to the tool developers' own personal judgments to create them. Yet because the numbers and percentages of a population that are placed within these risk bins matter to defendants, officials, and the public, these types of judgments demand direction, or oversight at the very least. The thresholds for risk bins are value judgments. As well, the choice of preferring false positives over false negatives (or vice versa) is political in nature.[446] Consequently, it should be the relevant policy makers, rather than developers, who make or drive such decisions. As an example, in Pennsylvania, the Adult Probation and Parole Department chose a 10:1 false positive to false negative ratio for their tool.[447] This significantly weighted valuation toward preferring false positives would likely be unpalatable to other stakeholders.

The threshold decision is a normative call. A value judgment that seeks to reduce false positives would set a higher threshold, while one that is more averse to false negatives would set a lower threshold.[448] These standards matter as they have real consequences to individuals. Thus, thresholds should be within the province of policy makers (or multidisciplinary implementation panel), not scientists.

The absence of industry norms on thresholds is reasonable. Thresholds are context-dependent. The (un)acceptability of certain false positive rates and false negative rates (or PPV and NPV) likewise will vary by context. A risk-informed decision on institutional placement holds a far different set of consequences than one regarding sentencing. A threshold set for adults may not be appropriate for juveniles. A jurisdiction with an appreciable array of well-funded social services and criminogenic needs-fulfilling programming may embrace a different threshold than one with few community resources.

Research to date has revealed that tools tend to generate a greater proportion of false positives than false negatives, resulting in risk-inflation practices.[449] It is not clear that agency officials in those jurisdictions are aware of this imbalance or whether they would approve if informed. The point here is to encourage a reform whereby the multidisciplinary panel can demand a public accounting of these thresholds and facilitate open debate on what they should be and how such choices serve legitimate interests.

It is important here to reflect on the purpose(s) for which the jurisdiction adopted a tool. If, for example, officials implemented an algorithmic scoring system in order to increase the rate of parole release, then adjusting the thresholds to place more individuals within the lower-risk bins would serve that goal more acutely than would a threshold with a higher percentage of individuals scored into the higher-risk bins. Similarly, if a sentencing system determines that it

[446] Sarah Valentine*, Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L. J. 364, 375 (2019).

[447] Garima Siwach & Shawn D. Bushway, *Adaption of Risk Tools to the Employment Context*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 475, 494 (Faye S. Taxman ed., 2017).

[448] Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 BEHAV. SCI. & L. 554, 559 (2018).

[449] Jodi L. Viljoen et al., *Do Risk Assessment Tools Help Manage and Reduce Risk of Violence and Reoffending? A Systematic Review*, 42 LAW & HUM. BEHAV. 181, 183 (2018).

wishes to use an algorithmic risk tool to reduce the rate of penalties involving a period of incarceration, then officials might intentionally opt to adjust the relevant thresholds of low, medium, and high accordingly. In contrast, a tool that is, instead, calibrated to prefer false positives may improperly limit the decision maker's discretion in considering lower-risk outcomes to screen out or divert.[450]

Policy Considerations:

The multidisciplinary team should address the placement of any thresholds, considering its goal(s), effects on predictive validity, individual fairness, and group fairness.

4. Communication

Many humans have problems understanding numbers. As previously discussed, the type of communication of risk assessment results can sway a factfinder one way or another. A jurisdiction should standardize a practice of risk communication so that individual evaluators and decision makers have a better and more consistent understanding of the meanings of the results conveyed. Participation by stakeholders could help the success of such standardization, such as ensuring that the language used is appropriate to that already in use within local agencies.

An operational decision needs to be made about how evaluators in the field should handle missing data. As some of the tools have become increasingly comprehensive, many data points must be collected. This also means multiple opportunities for evaluators to be unable to ascertain sufficient information to score needed factors. The question then is whether to score the individual anyway or issue no finding. Still, such a "no finding" must be addressed so that the decision makers do not perceive a negative connotation from it. Unfortunately, in at least one pretrial jurisdiction, the submission of "no recommendation" on release as a result of missing data was treated as a negative recommendation implying pretrial detention.[451]

Policy Considerations:

Risk communication practices ought to be standardized in an agency and/or jurisdiction. This might be done by the multidisciplinary implementation panel. Training on such standardization should be offered to evaluators and end users.

Risk communication should clarify the group-based nature of assessment practices.

Positive framing (as in the number or percentage of those who did not reoffend) may be preferable over negative framing.

The decision framework should address how to handle missing data.

[450] Nathan James, Cong. Res. Serv., *Risk and Needs Assessment in the Criminal Justice System* 14 (Oct. 13, 2015), https://digital.library.unt.edu/ark:/67531/metadc795663/m1/1/high_res_d/R44087_2015Oct13.pdf.

[451] Malcom M. Feeley, *How to Think About Criminal Court Reform*, 98 B.U. L. REV. 673, 693 (2018).

5. Overrides

Attention should be paid to any instructions or guidance that the developer issues on the use of overrides, either formally in codebooks or informally during implementation training sessions. The multidisciplinary panel may determine to reject or modify these dictates. Overrides do not present as scientific imperatives. They represent value-laden policy decisions.

The decision framework should therefore address how to consider policy and discretionary overrides. Any debate about overrides must also weigh them against the likelihood of reducing predictive ability. This does not mean that overrides are never acceptable. It could be that the goal(s) for implementation in the first place dictates an override to achieve that goal even at the expense of predictive ability.

Alternatively, local jurisdictions generally are afforded the ability to adjust their policies to confront unique community issues. As an example, a community overwhelmed with gun violence may well welcome an initiative to deny pretrial release to those arrested on gun charges. In such a case, officials may determine that the benefits to the public at large in the local area justify an override for gun arrestees. Or a city may be overburdened with too many of its young citizens being incarcerated pretrial, with its disproportionate effect on young minorities. A response there could be to engage a policy override for youths in a specified age range to lower risk to encourage a higher release rate.

In other words, risk assessment policies may need to be weighed against other legitimate concerns. Again, these are value judgments in which the scientific veneer of risk assessment does not justify removing the practice from real-world considerations.

Policy Considerations:

The multidisciplinary implementation panel should consider and give clear guidance on policy and professional overrides and the discrete justifications for them.

Discretionary overrides necessitate specific explanations in individual cases and should be subject to substantial oversight.

Agency administrations should keep track of overrides and regularly compare override rates between evaluators in order to improve consistency in assessment.

Risk communication to decision makers and to defendants must include transparency on whether an override was used, its form, why it was used, and overall rates of overrides. If an individual assessment is the result of an override, a statement should be required that overrides tend to reduce predictive ability of the tool.

B. Accountability

Importantly, no formal mechanism in the law or in the sciences exists to consistently enforce any form of algorithmic accountability.[452] There are several consequences to such a gap. For one,

[452] *See* Robyn Caplan et al., *Algorithmic Accountability: A Primer* 10 (Apr. 18, 2018), https://datasociety.net/pubs/alg_accountability.pdf.

some agencies are shielding their use of algorithmic tools from typical oversight by such methods as accepting privately funded tools or monies to engage in risk assessment.[453] Agencies may also cite nondisclosure agreements or tool owners' claims of trade secrets as a justification not to provide information on their predictive systems and practices.[454] Covert implementations such as these do not serve the interests of justice on behalf of the individuals who are targeted thereby.

Despite the many advantages of algorithmic assessment, risk profiling may fail to alleviate all the harms of mass incarceration as some "scholars are suspicious that contemporary extensions of risk assessment and risk reduction will likely only reproduce, or may even exacerbate, the injustices of contemporary criminal justice policy under a more 'objective' guise."[455] Thus, observers call for third-party auditing to engage in any form of scientific inquiry that may reveal information about the empirical validity and fairness of what are often black-box tools.[456] Such information will be useful to legal practitioners and policy makers in considering or reevaluating the use of algorithmic risk assessment to inform criminal justice decisions that carry significant consequences to individuals.[457]

1. Third-Party Audits

Risk tool developers tend to keep their algorithms and their data to themselves. Criminal justice agencies likewise are opting toward secrecy, despite legal obligations of transparency that otherwise are placed on them as governmental institutions. Any validation studies that either group affirmatively makes public is of value, of course. But such (often meager) publications are insufficient as there are many ways to manipulate the data (consciously or not) to serve their self-interests in promoting respect for, and deflecting criticism toward, their tools and practices. Such a stance improperly may hide to what extent and how the algorithm may produce systematic, individual, or group inequities.

The new FATML network is actively engaged in promoting methods for data science to ensure algorithms are used in ways that are seen as fair.

> Fairness can . . . be related to the notion of transparency—the question of how much we are entitled to know about any automated system that is used to make or inform a decision that affects us. Hiding the inner workings of an algorithm from public view might seem preferable, to avoid anyone gaming the system. But without transparency, how can decisions be probed and challenged?[458]

Three types of opacity are recognized: (a) intentional opacity, in which the tool owners declare trade secrets; (b) illiterate opacity, in which most people do not have the scientific skills

[453] Sarah Valentine*, Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L. J. 364, 376 (2019).

[454] Sarah Valentine*, Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L. J. 364, 377 (2019).

[455] Seth J. Prins & Adam Reich, *Can we Avoid Reductionism in Risk Reduction?*, 22 THEORETICAL CRIMINOLOGY 258, 259 (2018).

[456] *See Id*.

[457] Jennifer Skeem et al., *Gender, Risk Assessment, and Sanctioning*, 40 LAW & HUM. BEHAV. 580, 590 (2016).

[458] Sofia Olhede & Patrick Wolfe, *When Algorithms go Wrong, Who is Liable?*, 14 SIGNIFICANCE 8, 9 (2017).

to understand how the tools were created or operate; and (c) intrinsic opacity, in which machine learning methods are hard for even the technically inclined to interpret.[459]

Requiring agencies to make available their data sets to independent researchers is a critical step toward transparency and accountability.[460] This includes access to validation, revalidation, and cross-validation data.[461] Indeed, an expert argues that, with rare exception, a tool should not be implemented until it has been externally validated by independent parties.[462] Validity reports issued by those with financial incentives or other personal or professional conflicts should be viewed with caution,[463] if not rejected.

An audit team could benefit from scientific and legal expertise. Too often the lawyers are unable to fully understand the empirical issues, while the scientists are not necessarily cognizant of civil rights issues.[464] The legal representatives would additionally be attuned to privacy issues, laws of evidence, professional privilege protections, ethical issues on behalf of the evaluators, relevant statutory obligations, and human rights interests.

ℹ *Policy Considerations:*

Independent audits at regular intervals will serve interests in transparency and accountability. The body or agency adopting the risk assessment tool should ensure that appropriate funding is built in to be able to employ adequately trained and knowledgeable auditors.

Relying on individuals, groups, or companies that are aligned with the risk assessment tool (e.g., developers, authors, consultants, employees) is not an appropriate alternative to truly independent auditors.

An audit should include revalidating the tool on the populations for which it is scored, addressing algorithmic fairness measures, and conducting inter-rater reliability tests.

Adopting/implementing tools without trade secret protections is a crucial step toward transparency and accountability. Agencies can develop their own tools without resorting to claims of trade secrets or choose among the many available that do not claim to be proprietary.

[459] Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges*, 31 Phil. & Tech. 611, 619-20 (2018).

[460] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[461] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 09:29 start time) (on file with NACDL).

[462] Seena Fazel, *The Scientific Validity of Current Approaches to Violent and Criminal Risk Assessment, in* Predictive Sentencing: Normative and Empirical Perspectives 197, 206 (Jan W. de Keijser et al. eds. 2019).

[463] T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 Eur. Psychiatry 134, 135 (2017).

[464] Melissa Hamilton, *The Biased Algorithm*, 56 Am. Crim. L. Rev. 1553, 1559 (2019).

> *Criminal justice agencies (with multidisciplinary panel oversight) should proactively facilitate and cooperate with third-party audits by providing periodic access to data. These data sets should include individual-level data (i.e., individual offenders) with scoring information on predictive factors, outcomes (points, scores, risk bins), sociodemographic data, and recidivism data. Additional information that would be useful for auditors includes internal audit materials, training materials, codebooks, and user guides.*

Agencies may reasonably be concerned about potential problems with providing data on offenders. However, the recent social experiment with the Broward County data set has suggested fears may be overwrought. To date, there do not appear to be any consequences to individual privacy by third-party usage of the data set.

2. Adversarial Allegiance

Despite the guise of science and objectivity, algorithmic risk assessment is not free of adversarial bias. Algorithmic tools may be assumed to provide objective scoring mechanisms that yield consistent results across evaluators. Yet the potential that evaluators offer risk assessment results that favor the side that employed them is real. One report found repeated instances of adversarial bias whereby evaluators in individual cases scored the same tool higher or lower, consistent with the interests of the side, prosecutor or defense, respectively, that appointed them.[465] Other studies bolster this idea of adversarial allegiance by expert witnesses in legal cases when conveying risk tool results.[466]

In one salient example, independent researchers studied the use of the Static-99R tool in sex offender civil commitment trials. (This is the revised version of Static-99.) Static-99R offers two sets of norms (meaning experience tables of reoffending rates in the training samples). One norm set is for a preselected high-risk/needs group that includes a disproportionate number of offenders with severe mental disorders, significant needs, and considerable criminal histories. The other norm set applies to what is referred to as routine samples. At the same Static-99R scores, the preselected high-risk/needs group has significantly higher recidivism rates than the routine samples. Static-99R developers generally leave it to individual forensic evaluators to determine which set of norms to use in any individual case. These independent researchers found significant evidence of adversarial bias. Evaluators hired by the prosecution were significantly more likely to choose the preselected high-risk/needs norms (94%) than state agency evaluators (64%), who in turn were more likely to choose the high-risk/needs norms than experts employed by defense counsel (33%).[467] The discrepancy was large: The odds of prosecutorial evaluators

[465] Melissa Hamilton, *Public Safety, Individual Liberty, and Suspect Science: Future Dangerousness Assessments and Sex Offender Laws*, 83 TEMPLE L. REV. 697, 744-49 (2011), http://epubs.surrey.ac.uk/842345/1/83%20Temple%20L%20Rev%20697.pdf.

[466] Stephane M. Shepherd & Danny Sullivan, *Covert and Implicit Influences on the Interpretation of Violence Risk Instruments*, 24 PSYCHIATRY PSYCHOL. & L. 292, 297 (2017) (citing studies); Caroline Chevalier et al., *Statis-99R Reporting Practices in Sexually Violent Predator Cases: Does Norm Selection Reflect Adversarial Allegiance?*, 39 LAW & HUM. BEHAV. 209, 210-11 (2015) (citing cases).

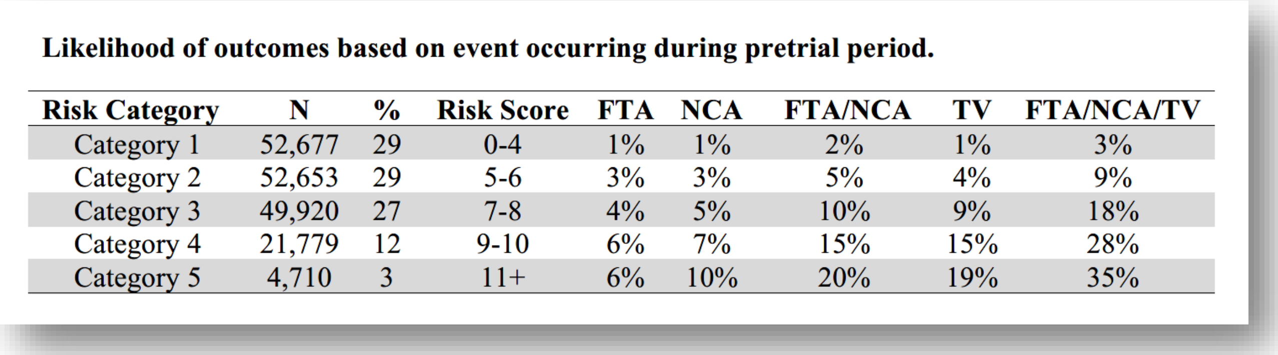
[467] Caroline Chevalier et al., *Statis-99R Reporting Practices in Sexually Violent Predator Cases: Does Norm Selection Reflect Adversarial Allegiance?*, 39 LAW & HUM. BEHAV. 209, 213 (2015).

choosing the high-risk/needs norms was 34 times that of defense evaluators.[468] Then when evaluators were asked to estimate the base rate of sexual recidivism for offenders they had assessed, prosecutorial evaluators on average estimated a significantly higher percentage than the other two groups, with defense evaluators estimating a lower percentage than state agency personnel.[469] At the same time, defense evaluators were far more likely than the other groups to be transparent about the tool's flaws by reporting accuracy statistics.[470]

A further form of adversarial bias exists if a side has the option of choosing which tool to utilize for a specific defendant. The widespread differences in them may present as encouraging a strategy akin to forum shopping. A prosecutor may select a tool with high base rates in the normative samples and/or a low threshold for high risk. On the other hand, a defendant may get lucky (or not) depending on whether the jurisdiction uses a tool that is "friendlier" to that individual's characteristics than another.[471]

An alternative example of an experience table may help illustrate multiple issues that have been so far addressed. Table 9 presents the experience table for the federal Pretrial Risk Assessment (PTRA) version 3.0.

Table 9: Example of an Experience Table



Likelihood of outcomes based on event occurring during pretrial period.

Risk Category	N	%	Risk Score	FTA	NCA	FTA/NCA	TV	FTA/NCA/TV
Category 1	52,677	29	0-4	1%	1%	2%	1%	3%
Category 2	52,653	29	5-6	3%	3%	5%	4%	9%
Category 3	49,920	27	7-8	4%	5%	10%	9%	18%
Category 4	21,779	12	9-10	6%	7%	15%	15%	28%
Category 5	4,710	3	11+	6%	10%	20%	19%	35%

Three categories of failure are provided: failure to appear (FTA), new criminal activity (NCA), and technical violations (TV). The PTRA includes five risk categories. One can observe that different definitions of failure are associated with significantly different rates of offending. Combining definitions obviously yields higher failure rates. This type of offering allows for evaluators to make choices in selecting failure measure(s) to report, which in turn will mean a potentially significant difference in the attributable risk outcome. At Category 5, an evaluator may choose between the FTA rate (6%), which is significantly lower than if the evaluator selects the combined FTA/NCA/TV failure rate (35%).

[468] *Id*. at 214.

[469] *Id*. at 214-15.

[470] *Id*. at 216.

[471] Angèle Christin et al., *Courts and Predictive Algorithms* 5 (Oct. 27, 2015), http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf.

Policy Considerations:

The decision framework should be clear when options are available (e.g., different experience tables, multiple recidivism measures), how to choose a particular option, and why to do so.

Counsel and end users should be cognizant of the potential for adversarial bias in evaluators.

3. Impact Assessments

The multidisciplinary implementation team should schedule and fund regular impact assessments. These exercises can help ensure that the tool is appropriately designed and applied toward the purpose(s) for which the tool was initiated.[472] An impact assessment could confirm, for example, the extent to which the results desired (e.g., *x*% reduction in pretrial release, ameliorating racial disproportionalities) have been achieved. An impact assessment may incorporate aspects of the third-party audit in terms of testing for systematic bias, individual fairness, and group fairness. Further, officials might be interested in discovering any unintended consequences. With this information, stakeholders could reevaluate and modify accordingly previous decisions on thresholds, assessment practices, and/or the framework.

The impact study might estimate how the tool may be affecting (positively or negatively) external decision-making. In a pretrial context, if judges are encouraged to use a tool to grant higher rates of release, it could be that police, prosecutors, or correctional staff alter their actions in expectation of the new release patterns. It is foreseeable here that a police officer who perceives that an offender is likely to be quickly released because of a potential low-risk score may decide not to arrest but instead to take another action that saves time and effort (e.g., issuing a notice to appear). In contrast, an officer who wishes to circumvent pretrial release of a particular offender may try to compensate by upgrading the arrest charge to a more serious offense or tacking on additional charges. Officials might in such a circumstance take whichever corrective actions may be seen as necessary (or not) to counteract such external environmental changes.

Policy Considerations:

Agencies should conduct regular impact assessments after a tool's implementation.

An impact assessment should include the following actions/elements: (a) an evaluation targeted to the goal(s) that the implementation of the tool was intended to address; (b) address the potential for external changes responding to the implementation; (c) consider intended and unintended consequences of the risk

[472] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 13:04 start time) (on file with NACDL).

assessment regime; and (d) evaluate how risk assessment practices impact due process rights.

4. Data Privacy

Scoring the tools often requires that evaluators collect sensitive data that then are input into the software system. As a result, confidentiality may be compromised. The information pertains to the offenders and potentially other people with whom they have associations. For example, information may be collected from public records, private files, and interviews. They entail such information as mental disorders, alcohol/drug problems, family, friends, relationships, and educational attainment.

Not only is there an issue of privacy, but such information may be used by police, prosecutors, or correctional officers as the basis of additional criminal charges or disciplinary actions. The previous discussion on offender interviews potentially eliciting self-incriminating information is relevant here.

In the future it might be anticipated that big data will increasingly be used to inform algorithmic risk assessments. A concern is that the use of big data in this way acts as a surveillance method.[473] Recent events have shown that big data profiling is largely unknown, untracked, and unregulated. It is foreseeable that one form of this surveillance will be for tool developers to engage in web scraping (data harvesting from websites) to exploit opportunities to collect data from websites (such as comments, photos, and videos posted on social media platforms used or accessed by defendants, their families, and friends). Web scraping theoretically could be used to score such existing predictors as drug use, criminal attitudes, and criminal associates. It also seems plausible to use web scraping to score on various criminal history measures in tools that do not require formal records to substantiate them. In this respect, presumably, web data may provide data to score criminal history events previously unknown to authorities but evidenced by social media information (e.g., illegal drug use; bragging about sexual, violent, or property offenses; removal of GPS trackers; violating curfew conditions).

Policy Considerations:

The implementation plan should include strong protections for data privacy.

5. Need for Lawyering

The foregoing policy comments apply more broadly to the choice of tools, assessment practices, and audits. At the same time, accountability can be enhanced at the individual case level. Due process protections are necessary. Tools are employed in decisions that have significant consequences to individuals in a criminal justice context. The guise of science via the

[473] Sarah Valentine*, Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L. J. 364, 370 (2019).

algorithms does not obviate the need for appropriate discovery and evidentiary hearings.[474] Indeed, algorithmic risk may justify an even greater occasion for open investigation in court.

Algorithmic outcomes are unlike normal evidence. Algorithms cannot directly be questioned or cross-examined.[475] Typically, potential evidence in criminal justice cases is not sheltered due to trade secret assertions. Plus, while corrections officers regularly engage in conversations with individuals they supervise, it is generally not in such a context as this, where the responses elicited are mechanistically documented and serve as a foundation for decisions that have significant consequences.

In an unpublished opinion, a state appellate court ruled, in a sentencing case in which an algorithmic risk score was included in a pre-sentence investigation report, that due process entitled the defense counsel to "the completed assessment in its entirety, including the questions asked/issues assessed, the answers provided/responses reported, and the numeric score assigned to the question or issue."[476] The court did not resolve whether counsel also had a right to the tool's confidential information as it recognized that no proprietary material had been sought at the sentencing hearing.[477]

In another case, a state supreme court judge recently warned of due process concerns whereby the general practice of giving a defendant the pre-sentence report information a few days before the sentencing hearing was insufficient time to study risk assessment information therein.

> But a few days' notice is not enough time for a defendant to mount a serious challenge to the underlying reliability of the risk assessment evidence as being so unreliable as to be hocus pocus. A full-court press on the question of reliability of the risk assessment would likely require the hiring of a highly qualified expert. Even if the defendant does not wish to mount a full-blown attack on the statistical model and instead wishes to make a more limited point—say, for instance, the disproportionate impact of use of housing, employment, and level of educational attainment of people of color—the defense will not be able to develop the attack in a few days, particularly when the defendant is indigent and will require court approval prior to the hiring of an expert to challenge the statistical information. And, of course, the state will want its opposing expert. In short, in order to allow the defendant to mount a substantial challenge to the underlying reliability of risk assessment data, and to give the state an appropriate opportunity to respond, the sentencing hearing will likely need to be continued for a period of weeks.[478]

[474] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[475] Science and Technology Committee, House of Commons, *Algorithms in Decision-Making* (HC 351) at 14, May 23, 2018.

[476] Kansas v. Walls, No. 116,027, slip op. at 6-7 (Kan. Ct. App. June 23, 2017).

[477] *Id*. at 7.

[478] State v. Guise, 921 N.W.2d 26, 34 (2018) (Appel, J. concurring).

And whether it be through a hearing or otherwise, it is beneficial to get as much information as possible about the risk tool and the individual's score into the record for purposes of appeal.[479] Any request for discovery should contain a demand for the identity, credentials, and relevant training of any evaluator who had input into or otherwise scored the risk tool.[480] It is possible that the evaluator failed to have sufficient education, training, and skills to have completed the particular instrument. The request might usefully include evidence of inter-rater reliability scores and the evaluator's record of compliance in terms of overrides.

Until a lawyer is well versed in the risk assessment arena, consultation is encouraged with knowledgeable data scientists and criminal justice researchers/academics in the field. Information on the methods used, statistical analyses, and potential empirical flaws will be enlightening and provide a foundation for challenges, if appropriate.

> [A]ssessment of the legal reliability of the use of data science, and computerized analytical processes to address a risk assessment requires (1) inquiry into the validity of the statistical analysis being run by the computers and (2) focus on the reliability of the multivariate data base being used.[481]

Of course, it is easy to suggest employing experts. It is more difficult in the field as the vast majority of defendants do not have access to funds to do so. In appropriate contexts, a part of the implementation plan could well be to require the agency adopting a tool to appoint one or more knowledgeable individuals to serve as something akin to court-appointed experts to explain the developmental process, decisions made (e.g., choosing predictors, thresholds, binning strategies), representativeness of sampling, etc. Having a knowledgeable human being explain the algorithm and how it affected the relevant decision fosters procedural due process.[482]

The algorithms and the questions within the tools are not as objective as might be assumed. Evidence of adversarial bias in scoring, as noted earlier, is substantial. Hence, unreflective faith in an evaluator's full independence is not recommended. Competent representation might require employing a sufficiently trained forensic evaluator to rescore the tool or employ what might be a more applicable tool considering the individual's sociodemographic profile and offense type. Key questions to raise about the science of algorithmic risk tools is available.[483]

[479] Transcript of Task Force Meeting, National Association of Defense Lawyers (Apr. 19, 2018, 14:59 start time) (on file with NACDL).

[480] John Philipsborn, *A Basic Assessment Toolbox: Aiming for Adequate Lawyering During the Spread of Risk Assessments*, CHAMPION 18, 18 (Jan./Feb. 2020).

[481] John Philipsborn, *A Basic Assessment Toolbox: Aiming for Adequate Lawyering During the Spread of Risk Assessments*, CHAMPION 18, 22 (Jan./Feb. 2020).

[482] Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. DAVIS L. REV. 1067, 1090 (2018).

[483] CROSS EXAMINING EXPERTS IN THE BEHAVIORAL SCIENCES (Terence W. Campbell & Demosthenes Lorandos eds., 2008-19), https://store.legal.thomsonreuters.com/law-products/Treatises/Cross-Examining-Experts-in-the-Behavioral-Sciences/p/102477862.

A resource of questions to ask experts about algorithmic risk tools generally, and with selected tools, is available.

A defense counsel's clientele will likely include both those who will benefit from low risk assessment and those who wish to avoid high-risk labels. Gaining expertise may allow one to argue for or against risk assessment in an individual case. And just because a lawyer argues for (or against) risk assessment in one case should not preclude one from taking a contrary position in another. Though some of the issues with risk assessment may be universal (all tools will have false positives and include factors that are proxies to sociodemographic characteristics), these tools are not fungible. Plus, each client's position with respect to the specific tool may be different. To illustrate, a lawyer in one case may appropriately argue that the VRAG is entirely suitable to his low-risk-scored client yet contend in another that VRAG is inappropriate because it was not properly validated on this latter client's population.

Policy Considerations:

Legal education groups should ramp up educational offerings regarding the law, science, policy, and ethics of all things risk assessment.

In individual cases, due process considerations mean that the defendant should have access to information on the design of the tool, validation studies, input factors, weighting, any thresholds for categorical bins, normative sample data, outputs, and override status.

Counsel may require more time to prepare for a hearing when an algorithmic risk score was involved in the decision.

6. The Right to a Human Explanation

The algorithmic turn has suggested, but certainly has made no strides in resolving, some new questions. To what extent do criminal defendants have a right for there to be substantial human involvement in a decision that infringes on such rights as freedom, privacy, and dignity? In other words, how much automation is acceptable? What is the extent of an individual's right to a human explanation for a decision that has significant consequences? Does the scientific guise of an algorithm alleviate due process concerns or exacerbate them because of nontransparency?

The position here is that there is some right to an interpretable explanation for why a decision was made. The extent of such right, just as with due process protections generally, depends on the context and importance of the decision to the individual. In the end, while algorithms may now be driving risk predictions, they inform on decisions that significantly affect human lives. The specter of the data scientist operating behind the proverbial curtain and sheltered by the veneer of science does a disservice to the transparency and accountability required from an already powerful criminal justice machinery. Policy proposals suggested herein are meant to bring the humanity back to the risk assessment model.

Policy Considerations:

Individuals have a right to a human explanation for a decision that has a substantial effect on their lives. An algorithmic risk score or categorical ranking cannot entirely replace human involvement in criminal justice decisions.

XII. GLOSSARY

Absolute risk	An estimate of the likelihood of recidivism given a score
Accuracy	The correctness of the tool's predictions
Actuarial tool	An instrument that includes factors that are empirically associated with recidivism that are weighted and combined into a total score
Algorithm	A computation that draws in inputs to process and then produces an output
Area under the curve	A statistical measure of the overall discriminative quality of a tool that conveys the probability that a randomly chosen recidivist would have been classified as higher risk than a randomly chosen non-recidivist
Balance for the negative class	Mean test score for recidivists are equivalent across groups
Balance for the positive class	Mean test score for non-recidivists are equivalent across groups
Base rate	The prevalence of the event in the relevant population (e.g., if 30% of the population reoffended, the base rate for that population is 30%)
Calibration	Absolute predictive accuracy, such as the extent of the correspondence between predicted rates and observed rates of recidivism
Classification error	Misclassifying a known outcome (such as assessing a recidivist as low risk)
Contingency table	A representation of a cross-tabulation between two variables, such as a risk prediction and the predicted outcome
Criminogenic	Likely to cause or result in crime (e.g., a system, place, or situation)
Demographic parity	The percentages of individuals predicted to recidivate (or otherwise at high risk) are equivalent across demographic groups
Desistance	The cessation of criminal or antisocial behavior by someone who has engaged in such behavior
Discrimination	Relative predictive accuracy, as in how well a tool distinguishes recidivists from non-recidivists
Dynamic factor	A predictive factor that is changeable
Equal calibration	The number predicted to recidivate is equivalent to the number of recidivists, and this is equivalent across groups

Error rate	The frequency that predictions of recidivist versus non-recidivist are incorrect
Evidence-based practice	A decision-making process that relies on the best available scientific research evidence
Experience table	A table that lists the observed rates of recidivism in the developmental samples, usually divided by risk level
False discovery rate	The reciprocal of the positive predictive value; the proportion of high-risk predictions who did not reoffend
False negative	An individual judged as low risk but who reoffended
False omission rate	The reciprocal of the negative predictive value; the proportion of low-risk predictions who reoffended
False positive	An individual judged as high risk but who did not reoffend
False positive rate	The reciprocal of the true negative rate; the proportion of those who did not reoffend who were predicted to reoffend
FATML	Fairness, accountability, and transparency in machine learning
Fitness	Relevance of the evidence to the legal purpose it is offered to inform
Follow-up period	The time frame in which individuals are tracked, usually after release into the community
Forecasting error	Incorrectly predicting the outcome of interest (such as assessing as high risk an individual who does not actually go on to recidivate)
Generalizability	The extent to which a study performed on a sample is applicable to other samples or to a larger population
Inter-rater reliability	A statistical measure of the extent to which two evaluators agree in their assessment results
Label bias	Mismeasurement of the outcome (e.g., recidivism)
Negative predictive value	The proportion of low-risk individuals who did not reoffend
Omitted variable bias	This occurs when a statistical model omits relevant causes of an outcome and signifies the difference between how much included variables actually affect the outcome and how much the model estimates that effect
Override	Substituting an actuarial or algorithmic risk outcome with another
Positive predictive value	The proportion of high-risk individuals who reoffended
Promotive factor	Characteristic that reduces the likelihood of recidivism and may predict desistance
Protective factor	Characteristic that can reduce the strength of a risk factor
Publication bias	Authors may simply not publish studies with insignificant findings

Recidivism	A negative criminal justice outcome that variously can be defined to include such events as rearrest, reconviction, supervision failure, technical violation, or return to confinement
Relative risk	A description of how one group's risk of recidivism compares with a reference group
Reliability	Consistency in results
Risk factorology	The ideology that algorithmic outcomes are driven by specified risk factors
Risk level	A type of numeric or ordinal ranking usually indicative of relative risk and typically greater levels signifying higher risk
Sample bias	Misrepresentation because of the use of a nonrepresentative sample
Static factor	Predictive factor that is unchangeable and typically historical
Statistical parity	The percentages of individuals predicted to recidivate (or otherwise at high risk) are equivalent across groups
Structured professional judgment	A method of risk assessment that involves scoring given items from a risk assessment tool and then using clinical judgment about additional factors in order to form an overall professional evaluation of risk
Test bias	A systematic error in how a test measures the members of a group compared with another group
Treatment equality	The ratio of errors (false positives/false negatives or false negatives/false positives) are equivalent across groups
True negative	An individual judged as low risk and who did not reoffend
True negative rate	The reciprocal of the false positive rate; the proportion of those who did not reoffend who were classified as lower risk
True positive	An individual judged as high risk and who reoffended
True positive rate	The proportion of recidivists who were classified as higher risk
Validation	The process of establishing the existence of evidence that the tool's methodology rises to an acceptable level of performance
Validity	The extent to which a test properly reflects the concept it is designed to reflect

For more information contact:

NATIONAL ASSOCIATION OF CRIMINAL DEFENSE LAWYERS©

1660 L Street NW, 12th Floor, Washington, DC 20036, Phone 202-872-8600

www.NACDL.org

NATIONAL
ASSOCIATION
OF CRIMINAL
DEFENSE LAWYERS

NACDL
FOUNDATION FOR
CRIMINAL
JUSTICE

This publication is available online at

www.NACDL.org/RiskAssessmentReport