U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

National Institute of Justice

Guidelines for Post-Sentencing Risk Assessment

Kristofer Bret Bucklen, Ph.D.
Pennsylvania Department of Corrections

Grant Duwe, Ph.D.
Minnesota Department of Corrections

Faye S. Taxman, Ph.D.
George Mason University

The authors are presented in alphabetical order to reflect the equal contributions to this paper by each party.

This paper was prepared with support from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, under contract number 2010F_10097. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily represent those of the Department of Justice.

NCJ 300654

NIJ | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

U.S. Department of Justice
Office of Justice Programs
810 Seventh St. N.W.
Washington, DC 20531

Jennifer Scherer, Ph.D.
Acting Director, National Institute of Justice

This and other publications and products of the National Institute of Justice can be found at:

National Institute of Justice
Strengthen Science • Advance Justice
nij.ojp.gov

Office of Justice Programs
Building Solutions • Supporting Communities • Advancing Justice
OJP.gov

The National Institute of Justice is the research, development, and evaluation agency of the U.S. Department of Justice. NIJ's mission is to advance scientific research, development, and evaluation to enhance the administration of justice and public safety.

The National Institute of Justice is a component of the Office of Justice Programs, which also includes the Bureau of Justice Assistance; the Bureau of Justice Statistics; the Office for Victims of Crime; the Office of Juvenile Justice and Delinquency Prevention; and the Office of Sex Offender Sentencing, Monitoring, Apprehending, Registering, and Tracking.

Acknowledgments

This paper would not be possible without the tremendous support of Dr. David Muhlhausen, former director of the National Institute of Justice, and Dr. Marie Garcia. They saw the need for a paper that would focus on moving the field forward in the area of risk and needs assessment. This paper is the result of these discussions. We appreciate their commitment to improving research and practice in the area of risk and needs assessment.

About This Paper

This paper is devoted to an important topic — risk and needs assessment (RNA) instruments for post-sentencing decision-making. It is primarily about risk assessment, but the principles also apply to needs assessments. In this paper, we discuss four principles that we believe are the most critical to guide both the design and implementation of data-informed decision-making tools: fairness, efficiency, effectiveness, and communication. Although there are other principles, our collective experience reveals that attention to these dimensions would benefit practitioners, researchers, industry leaders, and the general public. In particular, we believe that practitioners should understand the facets of quality RNA tools to ensure that their agency benefits from using state-of-the art design processes and implementing them to achieve high fidelity to the goals of proper prediction. Researchers will benefit from a visible set of guidelines to ensure that their partnerships with practitioners achieve the same goals of high-quality design and implementation.

Over the past several decades, the use of RNA in correctional systems has proliferated. Indeed, the vast majority of local, state, and federal correctional systems in the U.S. now use some type of RNA. These instruments are often used to allocate limited resources more effectively by calibrating the amount and type of programming to address the assessed risk and needs. More specifically, if an individual is assessed as high risk at the time of intake, she or he would not only be prioritized for programming that addresses the dynamic risk factors (criminogenic needs) but would also receive multiple assessments prior to release from prison or jail to determine whether and to what extent the risk has been reduced. Thus, the primary purpose of RNA assessments is to provide supportive and therapeutic resources to those who need them the most. Similarly, the tools can be used to ensure that the liberty restrictions (curfews, contacts with others, etc.) are appropriate to the risk that an individual presents.

Although RNA tools used for rehabilitation purposes are typically administered after an individual has been sentenced to probation or prison, these instruments have also been used prior to sentencing for pretrial detention decisions or, in some cases, to inform sentencing decisions. Post-sentencing RNA tools are more likely to adhere to rehabilitative principles compared to presentencing tools used when the person has not been convicted. Presentencing assessments are typically used to inform justice-related decisions, such as whether someone should be detained pretrial or released. Because the

delivery of therapeutic programming seldom occurs during the pretrial period, pretrial risk assessments do not tend to focus on treatment. Moreover, given that pretrial assessments tend to be administered only once, there is no subsequent opportunity for a person to demonstrate changes in their risk. Presentencing assessments are used for a different purpose than post-sentencing efforts, which is largely why they have attracted increased scrutiny and criticism with respect to fairness.

Using our four principles, we have outlined guidelines to help practitioners and researchers achieve the goals of improved RNA tools for the post-sentencing stage of the criminal justice system. The checklist provided in the conclusion of this paper will be useful to the field at large.

About the Authors

This paper is a product of the collective experiences and wisdom of three researchers in the field who are dedicated to the improvement of correctional practice. Each author brings their experience in the field as a researcher and/or practitioner, and each one has developed their own thoughts on how to improve the practice of corrections.

Kristofer "Bret" Bucklen, Ph.D., is Director of Planning, Research, and Statistics and Chief Scientific Officer for the Pennsylvania Department of Corrections. He has worked in a variety of positions in Pennsylvania's criminal justice system, including at the Pennsylvania Board of Probation and Parole, Pennsylvania State Police, and Pennsylvania Commission on Crime and Delinquency. He has extensive experience in designing, developing, and implementing risk assessment instruments, including designing Pennsylvania's risk screening tool and assisting in the development and implementation of Pennsylvania's Violence Forecast Model. Dr. Bucklen's recent work in the area of risk assessment has focused on using machine learning techniques and automation in order to improve predictive performance. His interest in the development of risk and needs assessment (RNA) is primarily in the actuarial aspects of risk assessment as applicable in a variety of criminal justice contexts.

Grant Duwe, Ph.D., is Research Director for the Minnesota Department of Corrections, where he forecasts the state's prison population, conducts program evaluations and research studies, and develops risk assessment instruments. He is the author of two books, and he has published more than 70 peer-reviewed articles on a wide range of topics in corrections. He has designed risk assessments for a variety of correctional populations that predict outcomes such as recidivism and prison misconduct. Dr. Duwe has developed the MnSOST-4 (Minnesota Sex Offender Screening Tool), MnSTARR (Minnesota Screening Tool Assessing Recidivism Risk), and MnSafeD (Minnesota Severe and Frequent Estimate for Discipline) for Minnesota's prison population, and he is a co-developer of the PATTERN (Prisoner Assessment Tool Targeting Estimated Risk and Needs) for the Federal Bureau of Prisons.

Faye S. Taxman, Ph.D., is a University Professor at the Schar School of Policy and Government and Director of the Center for Advancing Correctional Excellence (ACE!) at George Mason University with a passion for advancing social justice through quality programs, services, and tools. Her work on RNA began when she was a doctoral student studying with Dr. Don M. Gottfredson, a leading researcher in the area of sentencing guidelines and risk tools. Since her graduate studies, she has specialized in the design

and implementation of RNA along with studies to generate evidence or to examine the effectiveness of new systems and methods to reduce recidivism. Dr. Taxman's perspective on RNA is shaped by her research labs with correctional agencies and partnerships focused on implementation. She is the principal investigator on the Coordination and Translation Center for the National Institute on Drug Abuse's Justice Community Opioid Innovation Network and a study of 900 counties funded by the National Institute on Mental Health-funded evaluation of Stepping Up. She has published over 220 articles, is the author of six books, has edited a special edition of *Risk and Need Assessment: Theory and Practice* from the American Society of Criminology's Division of Corrections and Sentencing, and has designed translational toolkits.

Executive Summary

Risk and needs assessment (RNA) tools are used within corrections to prospectively identify those who have a greater risk of offending, violating laws or rules of prison or jail, and/or violating the conditions of community supervision. Correctional authorities use RNA instruments to guide a host of decisions that are, to a large extent, intended to enhance public safety and make better use of scarce resources. Despite the numerous ways in which RNA instruments can improve correctional policy and practice, the style and type of RNA currently used by much of the field has yet to live up to this promise because it is outdated, inefficient, and less effective than it should be.

In an effort to help the corrections field realize the potential that RNA instruments have for improving decision-making and reducing recidivism, we have drawn upon our collective wisdom and experience to identify four principles that are critical to the responsible and ethical use of RNAs. Within each principle is a set of guidelines that, when applied in practice, would help maximize the reliability and validity of RNA instruments. Because these guidelines comprise novel, evidence-based practices and procedures, the recommendations we propose in this paper are relatively innovative, at least for the field of corrections.

- The first principle, ***fairness,*** holds that RNA tools should be used to yield more equitable outcomes. When assessments are designed, efforts should be taken to eliminate or minimize potential sources of bias, which will mitigate racial and ethnic disparities. Preprocessing, in-processing, and post-processing adjustments are design strategies that can help minimize bias. Disparities can also be reduced through the way in which practitioners use RNAs, such as delivering more programming resources to those who need it the most (the risk principle). Collectively, this provides correctional agencies with a strategy for achieving better and more equitable outcomes.

- The second principle, ***efficiency,*** indicates that RNA instruments should rely on processes that promote reliability, expand assessment capacity, and do not burden staff resources. The vast majority of RNAs rely on time-consuming, cumbersome processes that mimic paper and pencil instruments; that is, they are forms to be completed and then manually scored by staff. The efficiency of RNA tools can be improved by adopting automated and computer-assisted scoring processes to increase reliability, validity, and assessment capacity. If RNA tools must be scored manually, then inter-rater reliability assessments must be carried out to ensure adequate consistency in scoring among staff.

- RNA instruments should not only be fair and efficient, but they should also be ***effective,*** which is the third key principle. The degree to which RNA instruments are effective depends largely on their predictive validity and how the tool is used within an agency. Machine learning algorithms often help increase predictive accuracy, although developers should test multiple algorithms to determine which one performs the best. RNA tools that are customized to the correctional population on which they are used will deliver better predictive performance.

- Finally, it is important to focus on the implementation and use of RNAs so that individuals can become increasingly aware of their risk factors. To this end, the fourth key principle is to employ strategies that improve ***risk communication.*** Training the correctional staff who will be using the RNA tool is essential for effective communication, particularly in how to explain the needs and translate it into a case plan. A risk communication system, which includes case plan improvement, treatment-matching algorithms, and graduated sanctions and incentives, provides an integrated model for decision-making that helps increase an individual's awareness of their own circumstances and need for programming.

These four principles are important for improving the transparency of RNA tools and providing fundamental guidelines to govern their development and implementation. Reliance on these principles can help RNA tools mitigate disparities and achieve better recidivism outcomes. Because this paper focuses on the design and utilization of RNA instruments, it is relevant for the developers of these tools and the practitioners who use them. To help developers and practitioners apply the guidelines outlined within each principle, we have provided a checklist at the end of this paper.

Introduction

Risk and needs assessment (RNA) generally involves predicting the likelihood of a negative outcome. Because RNA is used across a variety of disciplines, such as financial lending, insurance, health care, psychology, and criminology, the predicted outcomes range from default on a mortgage to patient mortality to recidivism. In predicting outcomes, RNA typically relies on algorithms, which can range from very simple to very complex. An algorithm used in an RNA instrument transforms the values for the items that predict the outcome into a predicted probability or risk score. The field of RNA has evolved over the past three decades given the sophistication of data, statistical methods, and technology. Based on this expansion, we believe that principles of effective design and implementation are needed to guide practitioners and researchers in their use of RNA. The principles we discuss in this paper are informed not only by our collective wisdom and experience, but also by the Risk-Need-Responsivity (RNR) framework, data science, and implementation science.

Within corrections, RNA is often used to prospectively identify those who have a greater risk of offending, violating laws or rules of prison or jail, and/or violating the conditions of community supervision. Correctional authorities use RNA to guide a host of decisions that are, to a large extent, intended to enhance public safety and make better use of scarce resources. Prior to the 1970s, risk assessment was based mostly on professional judgments made by staff, which left the system open to unbounded discretion. In the 1920s, objective, actuarial methods for assessing risk became available (Burgess, 1928). In the 1970s, professional judgment gave way to the emergence of actuarial-based tools, which Bonta and Andrews (2007) described as second-generation RNA instruments.

Evidence continues to accumulate that challenges the infamous conclusion drawn in the 1970s that "nothing works" (Martinson, 1974). The rise of the "what works" literature within corrections gradually led to the emergence of the principles of how to implement correctional interventions that improve outcomes for individuals in the justice system. The RNR model emerged as a set of principles that should guide implementation to deliver high performance-related outcomes. RNR provides a framework for determining (1) who should be treated (risk), (2) what areas should be addressed (needs), and (3) how to tailor the responses to individual factors that affect receptivity and performance in programming (responsivity) (Andrews, Bonta, and Wormith, 2006).

The RNR framework places a premium on prioritizing individuals for programming based on the use of valid and reliable RNA instruments. The RNA tool is central to contemporary correctional practice because it provides valid, objective information about the individual to inform decisions. The RNR paradigm assumes that interventions targeting criminogenic needs (dynamic risk factors) are more likely to decrease recidivism because individuals can make changes in their lives, decisions, and opinions to be crime (and drug) free. The emphasis on identifying criminogenic needs (dynamic risk factors) figured prominently in the development of RNA instruments where both static and dynamic factors are included. For nearly three decades, justice agencies have been encouraged to use RNA to standardize information that is used in justice decisions, reduce discrepancies and disparities in decisions made, properly use scarce justice resources, and develop a system that is focused on improving outcomes of justice-involved individuals based on science (Taxman, 2018a).

The RNR framework (Andrews and Bonta, 2010) is one of many sources for the principles outlined in this paper. Although the RNR model has become a familiar framework for implementing risk and needs assessment within the field of corrections, in a sense the principles of RNA outlined in this paper transcend (and are not limited to) the RNR framework. Risk and needs assessment can serve other purposes within the post-sentencing corrections environment, such as determining safe custody levels of inmates, informing parole release decisions and supervision levels, and choosing appropriate responses to technical parole violations.

RNA instruments in the criminal justice system have recently come under scrutiny due to concerns about the tools' bias and fairness. Much of this concern has been connected to presentencing risk assessments, which have been used to help determine whether an accused person in a criminal case should be confined or released (i.e., pretrial detention) or, for persons who have been convicted, the length of their criminal sentence. Given the disproportionate involvement of racial and ethnic minorities in the criminal justice system, combined with the fact that risk assessments rely (to a large extent) on historical data to make predictions about future behavior, some critics have argued that RNA instruments perpetuate or even exacerbate existing disparities (Angwin et al., 2016; Doleac and Stevenson, 2016). For example, if a person's criminal history is the by-product of prior discriminatory decisions and practices, even if only in part, then a risk assessment's predictions about future behavior may contain this bias. In addition, if the instrument is being used to confine higher-risk individuals for pretrial detention or give them longer sentences, then use of the tool might contribute to unfair outcomes where some people are treated more punitively due to their risk assessment score.

With post-sentencing RNA instruments, the goals are generally more rehabilitative; this is reflected in the emphasis placed on needs assessment and, in particular, the identification of dynamic risk factors that can be targeted for interventions. Apart from this consideration, however, there is a larger question worth raising: Is there a good alternative to data-driven RNA tools? Without actuarial-based risk predictions that standardize the information used to make a risk calculation, the only viable alternative would be for correctional staff to rely on their own professional judgment. The problem with professional judgment is the subjectivity that is introduced and the lack of methods to bound that subjectivity. Evidence from a variety of disciplines, including corrections and criminal justice, has consistently shown that actuarial RNAs

(i.e., statistically informed predictions) outperform clinical or professional judgment in terms of accuracy and reduce subjective decisions (Andrews and Bonta, 2010; Dawes, Faust, and Meehl, 1989; Duwe and Rocque, 2018; Wormith, Hogg, and Guzzo, 2012; Taxman, 2017).

Despite the numerous ways in which RNA instruments can enhance correctional policy and practice, the reality is they have yet to live up to their promise. The RNA tools in use tend to mimic "paper and pencil" instruments; that is, they are forms to be completed and then scored by correctional staff (either on the computer or on paper). When RNA tools are administered, we know — on the basis of nearly 50 years of work in this area and the collective experience of the authors of this paper — that staff do not always use the results from these assessments to create case plans, which facilitate the delivery of programming or help the individual understand his or her own risk factors. Nor do staff consistently communicate the results of these assessments to those on their caseloads. Overall, the style and type of RNA currently used by much of the field is outdated, inefficient, and substantially less effective than it should be.

In an effort to help RNA instruments realize their potential for improving correctional policy and practice, we have identified four principles that are critical to the responsible and ethical use of RNAs. Within each principle is a set of guidelines that, when applied in practice, would help maximize the reliability and validity of RNA instruments. Because these guidelines comprise novel, evidence-based practices and procedures, the recommendations we propose in this paper are relatively innovative, at least for the field of corrections.

- The first principle, ***fairness,*** holds that RNA tools should be used to yield more equitable outcomes. When assessments are designed, efforts should be taken to eliminate or minimize sources of bias, which will mitigate racial and ethnic disparities. Disparities can also be reduced through the way in which practitioners use RNAs, such as delivering more programming resources to those who need it the most (the risk principle). Transparency in the design and use of RNAs is a key component of fairness.

- The second principle, ***efficiency,*** indicates that RNA instruments should rely on processes that promote reliability, expand assessment capacity, and do not burden staff resources. The vast majority of RNAs rely on time-consuming, cumbersome processes that mimic paper and pencil instruments; that is, they are forms to be completed and then manually scored by staff. RNA tools offer efficiency in an overburdened system, including reducing unnecessary discretion in decision-making.

- RNA instruments should not only be fair and efficient, but they should also be ***effective,*** which is the third key principle. The degree to which RNA instruments are effective depends largely on their predictive validity and how the tool is used within an agency. A number of evidence-informed practices can be used in the development and validation of RNA tools that have been found to boost predictive performance. Similarly, a number of issues can adversely affect the effectiveness of the tool to achieve that performance in practice.

- Finally, although it is critical to use rigorous methods to design RNA instruments, it is just as important to focus on implementation and how the tools are used in practice. To this end, the fourth key principle is to employ strategies that improve ***risk communication.*** In particular, the paper identifies promising approaches for communicating the results from an RNA to correctional populations and staff.

As the use of RNAs for correctional populations has grown, so have misconceptions about how they are designed and used. Accordingly, we believe there should be industry guidelines that govern the responsible and ethical development and deployment of these instruments. The guidelines we present in this paper apply primarily to post-sentencing RNA tools. Although this paper focuses on the assessment of risk, the principles also apply to needs assessments given that similar issues of fairness, effectiveness, efficiency, and communications are applied to the various need domains (substance use, values, family, intergenerational issues, etc.). Given the extensive number of needs domains, discussing each of them is beyond the scope of this paper.

The following sections contain a review of the existing research underlying these four principles. Based on this review, we identify evidence-informed strategies that should, in our view, comprise industry guidelines for the development and use of RNAs. Finally, we conclude by summarizing the importance of these principles for correctional policy and practice by offering a checklist for developers/researchers and practitioners. Our innovative approach is based on principles that apply to both the design and implementation of RNA.

Principle One: Fairness

Recently, RNA instruments have been characterized as unfair or biased, usually regarding racial and/or gender groups (Pretrial Justice Institute, 2020). An important question that should be raised is: Fair as compared to what? The alternative to RNA instruments would be individuals relying on their own professional judgment to make assessments about what an individual may do in the future. Although humans typically bring a variety of biases to their decision-making, the goal is to incrementally reduce unaided biases. We believe that RNA instruments can be powerful tools for reform that help correctional systems achieve more equitable outcomes. This potential can be realized only if these instruments are properly designed and used in practice.

Achieving fairness in RNAs consists of two distinct components: (1) design issues and (2) use issues. Efforts to produce greater equity through the design of RNAs focus on the data and algorithms used to yield predictions that are free of disparities. This component assumes the onus for attaining fairness in RNAs falls largely on the developers of these instruments. When attention is given to how RNAs are used, the responsibility for producing greater equity lies more with the operational agencies (correctional systems) and, more narrowly, the practitioners who use these instruments to make decisions. This approach assumes that consistent application of the risk principle will help reduce disparities in outcomes.

A. The Case for Fair RNA Instruments: Design Issues

RNA tools are designed to improve decision-making in the justice system by ensuring that similar information is used in determining the risk that an individual poses to society. Subjective decision-making at the hands of line staff — whether officers, prosecutors, judges, or case managers — allows for an individual's preferences and perspectives to enter into a decision. Although decisions that rely on discretion may have some merits, they are subjective and based on the "eye of the beholder." With over 175 different types of cognitive biases that affect human decision-making, including anchoring, confirmation bias, group attribution error, fundamental attribution error, base rate fallacy, anecdotal fallacy, and telescoping effect (Benson, 2016), structured information can reduce biases.

Actuarial RNA tools remove discretion by focusing attention on key factors identified as being related to the outcome of interest (i.e., arrest, conviction, or incarceration). When staff use an instrument that standardizes the information that is collected and provides

an objective basis to sort and classify individuals, there is the potential to reduce disparities in how information is used. The resulting risk score reflects the probability of an individual engaging in negative behavior (or being successful). The risk score is often converted into a classification scheme with discrete categories that reflect the severity of the behavior.

On the surface, RNA tools appear to be objective and driven by the data. But fairness often comes from the methods that are used to create the instrument (methods) or the way in which individuals use the information (implementation). That is, the concept of fairness is rooted in the principles of equal treatment and equal outcomes for equivalent events/characteristics. Statistical predictions should minimize errors and be similar across groups regardless of demographic traits (Beretta et al., 2019; Berk et al., 2018; Breiman, 1996; Corbett-Davies et al., 2017; Dressler and Faird, 2018).

Statisticians have identified several core measures to examine the degree to which RNAs promote equal treatment and outcomes, as described in Table 1. These measures of fairness refer to predictive accuracy and could be used to determine whether an instrument distorts group differences. It is unlikely that all six standards can be simultaneously achieved because of the trade-offs between accuracy and bias reduction. Several challenges exist to achieve equality, including differential base rates among the groups, classification schemes, and different choices by different stakeholders. Unequal base rates among groups is typically the norm, which requires the use of different statistical methods to calibrate to overcome this unequal base rate.

Data Strategies To Improve Fairness

Researchers have identified three stages where methodological issues may affect the accuracy and fairness of an instrument, especially given the sources for the underlying data and/or variables used: preprocessing, in-processing, and post-processing issues (Romei and Ruggieri, 2013). *Preprocessing* requires assessing the source data for various types of biases that might exist due to how the data are collected, stored, measured, and generally reported. The goal of preprocessing activities is to remove any sources of unfairness in the data *prior to* the development of algorithms or risk calculations. Several preprocessing efforts can be used to address potential areas for bias: (1) the strength of each predictor, (2) how the

Table 1: Measures of Fairness

1.	Overall accuracy – Equal model accuracy between each class within a protected group, but does not distinguish between false positives or false negatives.
2.	Statistical parity – Equal marginal distributions of the predicted outcome for each class within a protected group (e.g., the fraction of black parolees forecasted to recidivate is equal to the fraction of white parolees forecasted to recidivate).
3.	Conditional procedure accuracy – Equal false negative rate and false positive rate between each class within a protected group (i.e., equal errors conditioned on the actual outcome).
4.	Conditional use accuracy – Equal positive predictive value and negative predictive value between each class within a protected group (i.e., equal errors conditions on the predicted outcome).
5.	Treatment equality – Equal ratio of false negatives to false positives between each class in a protected group.
6.	Total fairness – All of the above conditions are met simultaneously.

risk score and predictors differentiate among groups, (3) constructing values that predict outcomes to ensure that no information adversely affects any groups, (4) redistribution of the marginal distributions to ensure that the base rates are similar or comparable, (5) using different rules of association (either directly or indirectly) to ensure that the predictors are unbiased, and (6) examining conditional probabilities to ensure unbiased estimates, when possible. Specifically, preprocessing requires that careful attention be given to the data sources, how variables are constructed, what biases exist in the data, and how to adjust variables a priori to equalize base rates. The call is to know how different data elements might favor one class over another.

A specific example is the review of arrest history databases that often provide the source for many actuarial-based RNA tools. History of arrests, convictions, incarceration, and other criminal justice data are captured differently by state or federal agencies. Some states have a central data source while others require researchers to link files together. For example, some criminal record databases include any type of offense (incarcerable traffic offenses, misdemeanor offenses, citations, felony offenses, etc.) while others restrict the records to high-level misdemeanor and felony offenses. Thus, variables such as number (and type) of arrests will vary depending on the source. Another common data source issue is that some jurisdictions divert certain types of arrests and others process in lieu of the formal justice system; some are recorded, some are not. A preprocessing activity would be to document the type of data contained in a criminal record and then perhaps construct variables to examine how many arrests occurred for different types of events (incarcerable traffic, misdemeanor, felony, etc.). This process then allows one to assess the source of any bias by assessing how best to construct the variables.

In-processing efforts can further help reduce sources of unfairness that negatively affect different groups. In-processing refers to building adjustments in the algorithms and/or classification procedures to account for any biases that might occur, such as identifying potential variables where statistical differences occur among the groups and then adjusting to reduce potential areas for bias. The range of in-processing activities includes adjusting cut-points for key measures, recoding certain variables to equalize the outcomes among the protected classes, and adjusting the final algorithm to maximize fairness. Instrument developers can use a series of sensitivity analyses to ensure variables and resulting algorithm(s) are fair across various groups.

The last phase, *post-processing,* involves making adjustments to the algorithms after they have been created. These typically include adjustments to improve the performance of the tool, which may reduce the instrument's accuracy. Post-processing analyses attempt to ensure equal performance for protected groups by examining false positive and/or false negative rates, using constrained optimization approaches that address risk-specific thresholds, and making adjustments after reviewing the accuracy measures to improve the fit. In an effort to remove proxies that affect group biases, these procedures do not adjust the underlying variables but focus on cut-off points, allocation of subjects to different categories, and accuracy of prediction for each group.

B. Using RNA Instruments To Achieve More Equitable Outcomes

As discussed above, much of the recent concern over fairness and bias with RNAs has concentrated on the data and algorithms being used. But if we focus

strictly on the design of RNAs and do not examine how they are used, then we limit the likelihood of achieving more equitable outcomes. There are several limitations in particular that emphasize the data and algorithms used.

First, the existence of group disparities does not, ipso facto, mean the data are inaccurate and that the algorithms are biased. Second, as Berk and colleagues (2018) have demonstrated, it is not possible to simultaneously maximize accuracy and fairness when there is substantial base rate variation for the predicted outcome. In many ways, using an RNA is like holding up a mirror to criminal justice system policies, practices, and decisions (Mayson, 2019). When the mirror shows us the degree to which group disparities exist, we do not like what we see — and for good reason. Yet, the search for the holy grail of fair and accurate algorithms is like trying to swap out the mirror we are currently using in favor of one that shows us what we want to see, not what we really look like. In other words, the desire for an RNA that yields predictions free of disparities is like wanting to use a "skinny mirror." Third, the irony with the preference for a skinny mirror RNA is that while it may reflect how we wish to be seen by minimizing disparities in predicted risk, it would obscure the problem areas that may require more work and attention.

To illustrate, suppose we have two persons in prison who, if all things were equal, would have a similar recidivism risk. However, given that all things are seldom equal, the first person has a higher recidivism risk (per the assessment being used) because he has a longer criminal history that is attributable, at least in part, to having grown up in a disadvantaged, high-crime community that was subject to aggressive policing practices. Given that people released from prison often return to the same neighborhood from which they came, let us further assume he will go back to the same community while the second person will be released to a more desistance-friendly location that has many resources. To successfully desist from crime, the first person will likely need more resources while he is in prison than the second person.

However, with an assessment that aims to remove disparities in predicted risk, the first person would not be prioritized for programming any differently than the second person because their level of risk would be similar. Yet, as we also noted with this example, the first person will be returning to a community in which it may be more difficult to desist from crime. In this instance, use of the skinny mirror RNA would not be helpful for this person because it deprived him of access to resources that may have facilitated a successful transition from prison to the community. RNAs that accurately reflect reality could be powerful tools for reform, but only if we use them responsibly to enhance correctional decisions and practices.

Correctional agencies can reduce disparities and, in doing so, achieve greater fairness by focusing on the use of RNA instruments. With post-sentencing RNAs that attempt to follow the RNR model, adhering to the risk principle provides a strategy for achieving better outcomes, including a reduction in disparities. The risk principle identifies those who would benefit from programming resources since higher-risk individuals tend to have a need for a higher, more intensive dosage of programming to desist from crime. Therefore, if we see disparities in outcomes like recidivism and, by extension, recidivism risk, we should see disparities in program participation if we are adhering to the risk principle. For example, we can assume we have a group that comprises 25% of the prison population but 50% of those assessed as high risk. We can further assume there is an effective, intensive program that is typically reserved only

for higher-risk inmates. If we abide by the risk principle, we should see disparities in reverse for program participation, with this group making up approximately half of all participants (even though they account for only a quarter of all prisoners).

If properly developed and validated, an RNA instrument will accurately predict who poses a higher risk for reoffending and who does not, but it will not tell us what we should do with people who are higher risk or lower risk. This is why it is imperative to draw a distinction between how an RNA is designed versus how it gets used. If it accurately predicts recidivism but its use exacerbates existing disparities, then the problem lies with its use. The solution, then, is not to redesign the RNA but to change correctional policy and practice so that RNA instruments are being used responsibly to lower the risk for those who need programming the most, thereby enhancing public safety.

Implementation in correctional agencies can have an impact on how the RNA tools can improve fair decision-making. Practices that impact fairness include limiting RNA tools used at intake or reassessments, using offense as a criterion for program placement instead of criminogenic needs, failing to employ quality standards, and failing to develop policies and procedures that integrate RNA tools into practice. Related issues concern training staff on the meaning behind each element of the RNA tool and how to use RNA information in case planning, compliance management, and program placement. To avoid the skinny mirror, attention should be given to how the RNA tools are used in routine decisions with supportive policies and procedures that serve to enhance equal treatment.

Principle Two: Efficiency

RNA instruments generally rely on algorithms that convert the values for the predictive items — such as criminal history, demographic characteristics, dynamic risk factors, and/or program participation — into a risk score. The process in which the values for the items on an RNA instrument are populated has been referred to as the scoring method (Duwe and Rocque, 2017). The values for items can be entered manually, usually by correctional staff, or they can be populated through an automated process. The type of scoring method used has significant implications for the extent to which an RNA is reliable, valid, and efficient (Duwe and Rocque, 2019). Regardless of which scoring method is used, the items included on an RNA should be accessible to the public to promote transparency, increase confidence in the tools, and facilitate an understanding of the factors that are driving the risk scores.

When a manual scoring approach is used, differences in how staff score an RNA tool can be due to subjectivity of the items on the instrument, inadequate training, staff workload, the amount of time it takes to complete an assessment, and data entry errors (Duwe and Rocque, 2017). More broadly, inter-rater reliability (IRR) looks at the degree of agreement, or consistency, between raters in scoring an instrument. IRR has been recognized as a critical component to RNA, mainly because it can potentially affect how well an instrument can predict the outcome. After all, in order for a manually scored instrument to perform well in predicting an outcome, it must first be used consistently by raters who are scoring the instrument.

By standardizing the process in which items are scored, automated scoring methods eliminate inter-rater disagreement. This does not mean that an automated RNA is impervious to the problems associated with flawed data. For instance, if an automated process electronically pulls information from a database that was entered incorrectly, then this error would be reflected in the automated assessment. But this type of data entry error is also likely to be present in a manually scored RNA. By using a standardized scoring process, automation removes potential error from the assessment of risk. In doing so, automated scoring processes can help improve the reliability and, thus, the predictive validity of RNA decisions (Duwe and Rocque, 2017).

Although there has been sparse research examining the impact the scoring method has on RNA instruments, there are several broad conclusions that can be drawn from the few existing studies that have been done. First, even a relatively modest amount of inter-rater disagreement can have a significant impact on predictive performance. In their study that examined assessment data from the MnSTARR (Minnesota Screening Tool Assessing

Recidivism Risk), a manually scored instrument the Minnesota Department of Corrections (MnDOC) developed and began using in 2013 (Duwe, 2014), Duwe and Rocque (2017) compared the reliability of a manual scoring approach with a fully automated process. Using multiple performance metrics, Duwe and Rocque (2017) then evaluated the predictive validity of the two scoring methods — manual and automated — across male and female prisoners for four measures of recidivism.

The results showed the MnSTARR was scored with a relatively high degree of consistency by MnDOC staff. Indeed, the intraclass correlation coefficient (ICC) values, which ranged from 0.81 to 0.94, are considered "excellent" (Hallgren, 2012). Duwe and Rocque (2017) reported the automated assessments significantly outperformed those that had been scored manually in predicting recidivism. They found that as inter-rater disagreement increased (i.e., the ICC value decreased), predictive performance significantly decreased. By ensuring that everyone is scored the same way, automated scoring methods eliminate the inter-rater disagreement that is inherent in manually scored assessments. In doing so, automated scoring processes can help improve the reliability and, by extension, the predictive performance of RNA tools (Duwe and Rocque, 2017).

Second, the increased consistency offered by automated scoring methods may also help mitigate disparities often observed in predicted risk, or risk levels. In a more recent study, Duwe and Rocque (2019) externally validated the MnSTARR on a sample of 3,985 inmates released from Minnesota prisons in 2014. Although the manually scored MnSTARR achieved adequate predictive validity, its performance would have been better with an automated scoring process. Just as important, Duwe and Rocque (2019) reported that although the MnSTARR performed better for whites than nonwhites (Black, American Indian, and Asian), the magnitude of this difference would have been minimized using automated scoring.

Third, and perhaps most important, the use of an automated scoring process has major implications for assessment capacity. Duwe and Rocque (2019) reported that only 52% of the 7,657 releases from Minnesota prisons in 2014 had been manually assessed on the MnSTARR for recidivism risk, and most received only one assessment. In 2016, the MnDOC implemented the MnSTARR 2.0, a fully automated, gender-specific instrument that assesses risk for multiple types of recidivism. More specifically, the MnSTARR 2.0 extracts data from the state's criminal history repository to populate the criminal history items and data from the MnDOC's management information system to populate items pertaining to demographic characteristics (e.g., gender, age, and marital status), institutional behavior (e.g., discipline convictions and gang affiliation), and participation in programming (e.g., earning a post-secondary degree in prison, completing chemical dependency treatment, and completing cognitive-behavioral therapy). Although the original MnSTARR took staff 35 minutes on average to score by hand, scoring the MnSTARR 2.0 does not require any additional staff time. As a result of a more efficient scoring process, every individual released from Minnesota prisons since 2016 has been assessed at least once and, in most instances, multiple times prior to release. During the first year alone, a total of 41,253 MnSTARR 2.0 assessments were completed. With the manually scored MnSTARR, it would have taken more than 24,000 hours in staff time (nearly the equivalent of 12 full-time employees) to score that many assessments. By saving that many staff hours, automating the MnSTARR 2.0 produced a cost-benefit estimate of $955,990 during its first year, resulting in a return on investment of $8.08 (Duwe and Rocque, 2019).

The available evidence suggests there are several implications for improving the efficiency of RNA instruments. First, given the advantages associated with automation, correctional systems should invest more resources in automating the scoring process. Automation can significantly increase the efficiency of the RNA process by eliminating the time that prison staff spend in manually scoring assessments and undergoing the training required for those who use the instrument. Even though automating the RNA process entails a cost for prison systems, it still delivers a highly favorable return on investment due to the significant increase in efficiency.

If auto-scoring is not feasible, then the use of technology, such as computer-assisted survey software, should be considered. With some assessments, especially those that also assess for needs, it may be necessary to collect the input data through either a survey or an interview with the probationer, prisoner, or parolee. Using computer-assisted survey software would significantly increase the efficiency of the scoring process. Rather than relying on staff to administer the assessments through a face-to-face interview, people in custody or under supervision should complete it on their own through a device such as a tablet or kiosk.

Finally, if an instrument must be scored manually, then it is necessary to demonstrate that it can be scored consistently. If the data must be manually entered by correctional staff, then an IRR assessment should be completed to determine the degree to which there is inter-rater disagreement among the staff scoring the assessment. Duwe and Rocque (2017) proposed the following ICC thresholds for assessing IRR within the context of manually scored RNA tools: 0.95 and above is excellent, 0.85 to 0.94 is good, 0.75 to 0.84 is adequate, and below 0.75 is poor.

Wave 3 (the most recent development) relies on machine learning algorithms to develop actuarial RNA instruments. Machine learning is a subset of artificial intelligence in which a model proceeds adaptively from the data through a process of training. Machine learning approaches differ from earlier statistical approaches in that they are not based on a parametric model that is imposed in advance on the data. Instead, the data itself inductively determine the structure of the risk model. Machine learning is a broad field that includes a host of different families of algorithms and subapproaches, such as classification and regression trees, k-means clustering methods, Bayesian networks, artificial neural networks, support vector machines, and ensemble methods like random forests and stochastic gradient boosting.

B. Machine Learning Methods

Machine learning approaches have been rapidly and widely adopted in the private sector for many types of predictive analytic applications by organizations such as Google, Microsoft, and Amazon. Formal proofs, simulations, and comparisons across many different datasets have generally demonstrated that machine learning approaches improve predictive accuracy above and beyond earlier parametric statistical approaches such as logistic regression (Breiman, 1996, 2001; Breitenbach et al., 2009; Chipman, George, and McCulloch, 2010; Duwe and Kim, 2015, 2016; Friedman, 2002; Hamilton et al., 2015; Hess and Turner, 2013; Vapnick, 1998). Recent studies, however, find that machine learning algorithms in criminal justice perform no better than older and simpler statistical approaches (Liu et al., 2011; Tollenaar and Van der Heijden, 2013). The weight of the evidence, though, suggests that forecasting accuracy will depend on the complexity of the forecasting situation, and that machine learning algorithms are superior when individual items work together in complex ways to predict risk (i.e., the complexity of the "decision boundary") and when the risk being predicted is for a relatively rare event such as violent recidivism (Berk and Bleich, 2013) or sex offense recidivism (Duwe, 2017).

One way of describing this advantage of machine learning is that it "squeezes more juice" out of risk factors to produce a risk score. Various machine learning algorithms should at least be tested side-by-side with conventional statistical approaches when developing an RNA instrument in order to determine which approach works best in any particular application. Exclusively relying on older (e.g., regression-based) statistical approaches is too limiting and will likely produce lower predictive accuracy than could maximally be achieved in many criminal justice applications.

To illustrate how machine learning can generate superior performance on an RNA tool, the concept of a decision boundary is important. The goal of a decision boundary is to differentiate two or more risk groups. For example, the decision boundary might be to differentiate between recidivists and nonrecidivists. Figures 1 and 2 show a decision boundary where only two items (age and prior criminal history) are included on an assessment tool. The scatterplots show the person's number of prior arrests on the y-axis and the person's current age on the x-axis. Red dots represent recidivists and blue dots represent nonrecidivists. The goal of developing a good RNA model here is to draw a line between the two-dimensional data that best separates recidivists (red dots) from nonrecidivists (blue dots). In Figure 1, this is fairly easy to do with a straight line (see the straight line that goes from the bottom left up to the top right of this figure). Figure 2 represents a more complicated relationship. Drawing a straight line through the data in Figure 2 would be suboptimal for distinguishing recidivists

Figure 1: Simple Decision Boundary

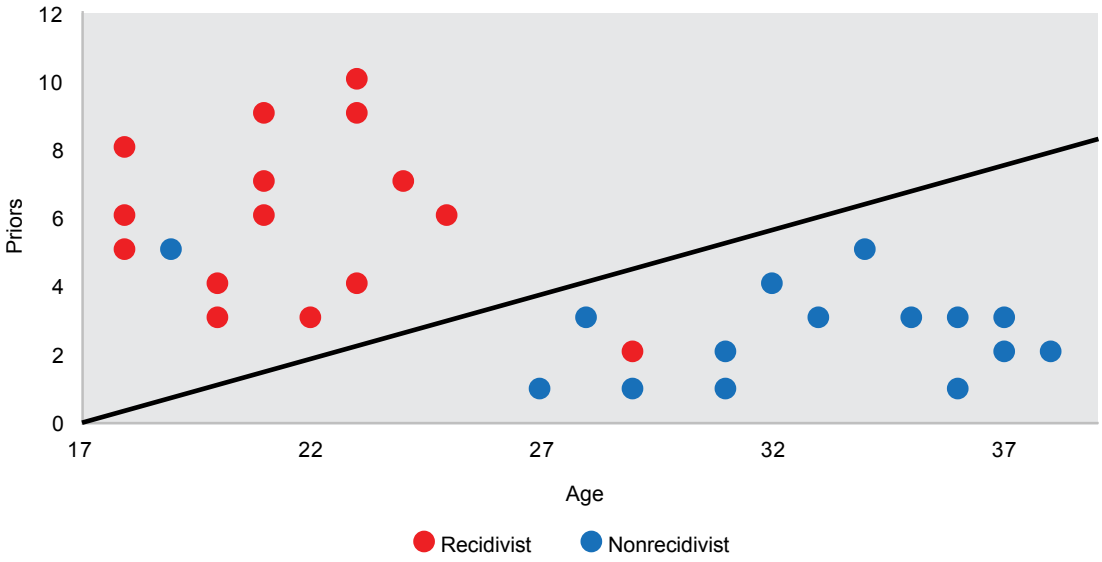
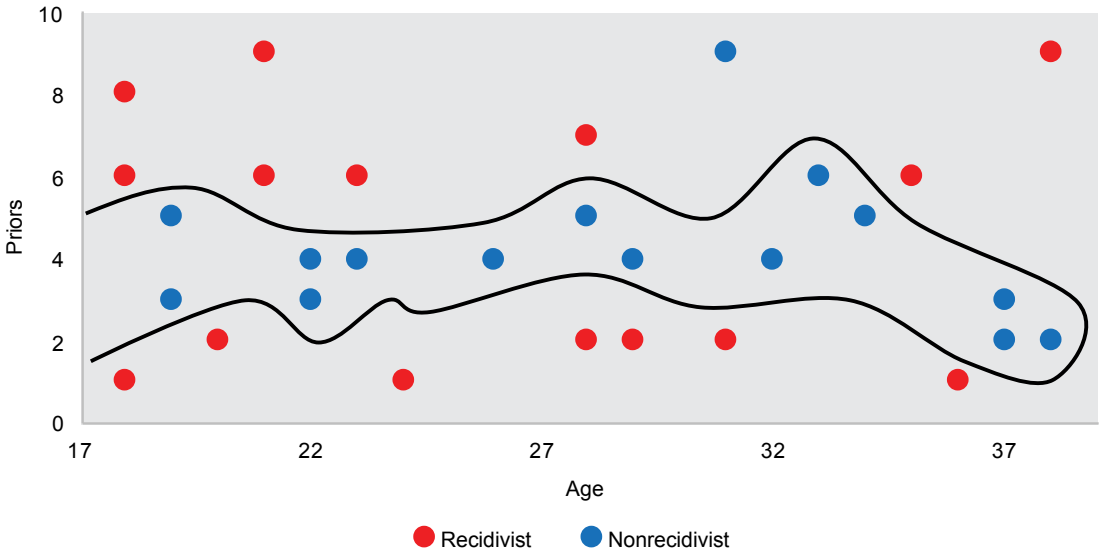


Figure 2: Complex Decision Boundary



from nonrecidivists. Instead, the wavy line represents a better split of recidivists and nonrecidivists. Now imagine that the RNA tool has 20 items rather than two, and the 20 items relate to each other in a complicated (nonlinear) way. This is exactly the type of problem for which machine learning has an advantage. Machine learning is better suited for complex decision boundary problems. Real life prediction problems often face the type of complexity that machine learning is well-suited to address.

C. Quantifying Predictive Validity

There are several statistical measures that can be used for establishing predictive accuracy. Table 2 provides a summary of some common predictive accuracy statistics.

At a basic level, predictive accuracy can (and should) first be examined using a classification table (or confusion table). In the simple example of forecasting recidivists versus nonrecidivists, the classification table takes the form of a two-by-two frequency table (Table 3).

This section discusses only the two most frequently used predictive accuracy measures from Table 2, given that the other measures are less commonly used. The most frequently used measure is the Receiver Operating Characteristic Area Under the Curve, or AUC. The AUC technically quantifies the discriminant accuracy of an RNA instrument by representing the trade-off between true positives and false positives at different thresholds of the risk score. The second most frequently used measure is the basic accuracy statistic (ACC). The ACC can be generated easily from a classification table (see Table 3). One advantage of the AUC over the ACC, however, is that the AUC is not a function of a pre-defined score cut-point for sorting forecasted recidivists from nonrecidivists. The AUC is independent of base rates and selection ratios, and thus is often the preferred predictive accuracy measure.

An AUC value ranges from 0 to 1, with the worst possible score being 0.5 and the best possible score being 1 or 0. One practical way to interpret an AUC score is the percent of the time that a recidivist scores higher than a nonrecidivist on an RNA instrument (assuming a higher score means a higher likelihood of recidivism). An AUC score of 0.5 means that 50% of the time that we compare a random recidivist and nonrecidivist, the recidivist scores higher than the nonrecidivist. This is no better than flipping a coin. On the other hand, an AUC score of 1 means that 100% of the time we compare a random recidivist and nonrecidivist, the recidivist scores higher than the nonrecidivist. This would be perfect predictive accuracy.

Standards for an acceptable AUC score for an RNA instrument are changing, in part as new methods such as machine learning algorithms are better able to achieve

Table 2: Predictive Accuracy Measures

Statistic	Name	Description	Interpretation
ACC	Accuracy	Measure of predictive accuracy that provides percentage of correct classifications	Values range from 0 to 1 (or 0 to 100%), with higher values indicating greater accuracy
AUC	Area Under the Curve	Measure of predictive discrimination	Values range from 0 to 1, with values at either end representing better discrimination
H	H-Measure	Measure of predictive discrimination developed by Hand (2009)	Values range from 0 to 1, with higher values indicating greater discrimination
ICC	Intraclass Correlation Coefficient	Commonly used statistic to assess inter-rater reliability	Values range from 0 to 1, with higher scores indicating greater reliability
RMSE	Root Mean Square Error	Measure of calibration between observed and predicted values	Values range from 0 to 1, with lower values representing better calibration
SAR	Squared Error, Accuracy, and Receiver Operating Characteristic	Composite measure of predictive performance developed by Caruana et al. (2004)	Values range from 0 to 1, with higher values indicating better predictive performance

Adapted from Duwe and Rocque, 2017.

higher predictive accuracy (i.e., higher AUC scores). One convention for RNA instruments that has often been referenced is that an AUC score between 0.64 and 0.71 is moderately predictive, and an AUC score above 0.71 is highly predictive (Rice and Harris, 2005).

A recent summary of the predictive validity of several popular RNA instruments for which AUC scores are published suggests that these tools average an AUC score of 0.68 (U.S. Department of Justice, 2019). However, more recent and improved instruments are consistently producing AUC scores well above 0.7. Instruments with scores in the range of 0.65 to 0.7 were once considered acceptable, but may become unacceptable as these new standards are adopted in the future. In addition, there are other objectives to be balanced when creating an RNA instrument, such as transparency, simplicity (or parsimony), and fairness. From a pure effectiveness standpoint, the full range of modern algorithms should be considered in order to build an instrument that maximizes predictive validity using a common metric such as the AUC.

Table 3 provides an example of a two-by-two classification table. The columns indicate the forecasted result (forecasted recidivists and forecasted nonrecidivists), and the rows indicate the actual result based on a training or testing dataset (actual recidivists and actual nonrecidivists). Any forecast model presents two types of error: false negatives and false positives. False negatives (b) represent the number of actual recidivists who are incorrectly forecasted to be nonrecidivists. False positives (c) represent the number of actual nonrecidivists who are incorrectly forecasted to be recidivists. The overall accuracy rate (ACC) of a model is simply the sum of the number of true positives (a) and true negatives (d) divided by the grand total of the number of cases in all four cells (a+b+c+d). Conversely, the overall error rate of a model is the sum of the number of false negatives (b) and false positives (c) divided by the grand total of the number of cases in all four cells (a+b+c+d), or simply 1 – ACC.

In most practical RNA applications, the cost of making a false negative error and the cost of making a false positive error are not equal. For example, a jurisdiction that is using an RNA instrument to inform parole release decision-making might be much more concerned with releasing on parole an inmate who was inaccurately forecasted to be a nonrecidivist (a false negative), than they are concerned with not releasing to parole an individual who was inaccurately forecasted to be a recidivist (a false positive). Each type of error has different implications. In the parole example, a false negative could mean that a person is released to parole who subsequently reoffends, jeopardizing public safety. On the other hand, a false positive could mean that a person's liberty is unfairly intruded upon by denying the person release to parole on the basis that he or she was inaccurately forecasted to be a recidivist. The benefit of exploring a classification table as a first step in examining an RNA instrument's predictive validity is that it illustrates each type of error separately, whereas other predictive accuracy statistics like the AUC only provide an overall number that combines the impact of false positives and false negatives together.

Table 3: Example Classification Table

	Forecasted Recidivist	Forecasted Nonrecidivist
Actual Recidivist	True positives (a)	False negatives (b)
Actual Nonrecidivist	False positives (c)	True negatives (d)

Table 4: Classification Measures

Measure	Calculation
Sensitivity (or Recall)	a/(a+b)
Specificity	d/(c+d)
False Negative Rate	b/(a+b)
False Positive Rate	c/(c+d)
Positive Predictive Value	a/(a+c)
Negative Predictive Value	d/(b+d)

Using the example classification table (Table 3), there are several quantities that can be generated for measuring each type of error in a model. These measures include Sensitivity (or Recall), Specificity, the False Negative Rate, the False Positive Rate, the Positive Predictive Value, and the Negative Predictive Value. Table 4 provides the calculation formula for each of these quantities, referencing the example classification table illustrated in Table 3.

In any RNA instrument, a trade-off between the false positive rate and the false negative rate is inevitable. It is impossible to decrease one error rate without increasing the other error rate. This is the simple mathematical reality of any RNA instrument. Administrators and policymakers thus must determine what level of each type of error is acceptable. There is no standard acceptable false positive rate or false negative rate. Determining the acceptable trade-off between these two error rates is a policy decision that is outside of the science of RNA development.[1] An RNA instrument will present administrators and policymakers with a "loss function" related to false positives and also a loss function related to false negatives. Since these two loss functions are rarely equal (i.e., the cost of each type of error is rarely valued the same), an RNA instrument presents a situation in which there is an asymmetric loss function. Another advantage of using newer machine learning algorithms is that the exact asymmetric loss function desired by policymakers can be built into the model up front.

D. Validation and Localization

Another principle of effectiveness that is a part of best practices and new guidelines in RNA instrument development is creating a localized and customized instrument. For much of the history of the use of RNA instruments in correctional settings, off-the-shelf RNA instruments were adopted. Many off-the-shelf instruments are proprietary and still widely used, such as the Level of Service Inventory-Revised (LSI-R), Correctional Offender Management Profiling for Alternative Sanctions, and Ohio Risk Assessment System. The unique aspects of different jurisdictions and the populations they serve can impact predictive ability. At a minimum, an off-the-shelf tool should be validated locally using local, jurisdiction-specific data before being adopted.

[1]In practice, the cost of each type of error resulting from the administration of an RNA instrument is usually valued differently. One advantage of most machine learning algorithms is that they can build cost trade-offs directly into the RNA model up front. More conventional statistical approaches such as logistic regression can only handle trade-off costs on the back end by adjusting score thresholds, and even then it can be difficult to attain the exact desired trade-off if the instrument does not contain a significant number of points in order to fine-tune the thresholds. In mathematical optimization terms, the quantification of how much a specific type of error in an RNA model costs practically is called a loss function (or cost function).

Although there may be some advantages to an off-the-shelf tool, there are also many disadvantages. RNA instruments are built to predict certain outcomes (usually recidivism), but it is also possible to design instruments to predict pretrial release, parole release decisions, housing classification decisions, and so on. A common belief is that versatile RNA instruments can be used for a number of criminal justice decisions, regardless of how they are developed. A generic RNA tool does not exist for every decision in the justice system, and users should ensure that the potential use is consistent with the tool's design.

Due to the limitations of off-the-shelf instruments, jurisdictions may find it better to create, validate, and revalidate an RNA instrument locally as a best practice for improving effectiveness. The field is starting to realize that off-the-shelf instruments do not work equally well in all settings or applications. For example, the Pennsylvania Department of Corrections (PA DOC) historically used the LSI-R as its primary RNA instrument until a validation study of the LSI-R in 2003 questioned its predictive ability as used in the PA DOC. The study found that the overall LSI-R score was not strongly predictive of recidivism using a large validation sample, and that only a small subset of items on the LSI-R actually predicted recidivism in Pennsylvania (Austin et al., 2003). This general finding of a lack of predictive ability among PA DOC inmates was again confirmed in another study a few years later (Simourd, 2006).

In a subsequent study, a new instrument was developed using Pennsylvania data from the LSI-R (Bucklen, 2007). This new instrument (later named the Risk Screening Tool) used the six most predictive items from the LSI-R based on Pennsylvania data and also added current age, which was not included on the LSI-R. This new tool was then compared to the LSI-R for predictive accuracy using a new sample of PA DOC inmates. It was found to produce superior predictive accuracy, increasing the AUC from 0.64 to 0.67 (a statistically significant increase) using only seven items rather than the 54 items on the LSI-R. Although many items on the LSI-R may have been predictive of recidivism in other jurisdictions or with other populations, they were not predictive of recidivism in the PA DOC and were thus creating "noise" in the assessment, which diminished its predictive ability. Due to this increase in effectiveness, the PA DOC later adopted the new homegrown tool (the Risk Screening Tool) in place of the LSI-R.

In addition to the PA DOC, other jurisdictions in the U.S. have designed and implemented home-grown assessment instruments. In a study comparing an off-the-shelf RNA instrument (the Static-99R) to a home-grown RNA instrument developed locally for assessing sex offender risk in Minnesota (the MnSOST-3), the authors found that the home-grown MnSOST-3 outperformed the off-the-shelf Static-99R on both measures of sex offender recidivism (Duwe and Rocque, 2018). To use a sports analogy, home-grown tools tend to have a "home-field advantage," which leads them to generally perform better than off-the-shelf tools.

Another example of the home-field advantage principle comes from the Federal Bureau of Prisons (BOP). Under the First Step Act, BOP was mandated to adopt an RNA instrument that was valid for its own population. The result of this mandate was the creation of an RNA instrument called the PATTERN. The authors of the report made a clear case for why it was not advisable for BOP to simply adopt an off-the-shelf RNA instrument from another jurisdiction (U.S. Department of Justice, 2019). Most importantly for this context, many off-the-shelf RNA tools were created or validated using state prison populations or correctional populations outside the United States. State prison inmates typically have more variety in their criminal history

than federal inmates, including a greater prevalence of violent histories. Since the nature and frequency of criminal history is almost always found to be an important predictor in an RNA instrument, this difference between state and federal inmates could have important implications for the predictive accuracy of an RNA instrument if BOP would have simply adopted a tool primarily validated on state prison inmates.

E. Big Data

Another consideration for improving effectiveness in RNA validation is the use of big data. With advances in computing power, as much data as available should be considered in constructing RNA instruments. The process of developing a highly predictive RNA instrument can be a largely atheoretical exercise. For example, if shoe size turns out to be a significant predictor of recidivism, even if one cannot come up with a good theory of why, then shoe size could be considered for inclusion on the instrument. The science of RNA development is an actuarial process that does not concern itself with causality.[2] As discussed in the Principle Two: Efficiency section of this paper, not only should as much data as feasible and available be considered, but special focus should be placed on making use of automated data from existing administrative databases. Preprocessing and in-processing activities should be used to assess the data.

Many agencies house a host of data elements in their administrative databases that they may or may not, on the surface, consider as being relevant to RNA. Modern computing and machine learning algorithms are well-suited to handle many data elements and very large sample sizes. A jurisdiction might consider cross-linking with data from other agencies to pull into an RNA instrument. This may be challenging in terms of data-sharing agreements and information technology infrastructure, but it may also significantly enhance the predictive ability of an RNA instrument. Cross-agency data sharing is becoming a frequent best practice for multiple uses in the public sector, with public health providing a good example (U.S. Department of Health and Human Services, 2011). Machine learning algorithms impose no inherent penalty on including more data elements; the only limiting factor might be practical implications for incorporating the data.

[2]One caveat is that developers of an RNA instrument should still consider the factors outlined in the Principle One: Fairness section of this report in regard to which factors may introduce bias even if they are significant predictors. For example, race may turn out to be a significant predictor of recidivism but could also introduce bias that should lead the developer to remove it as an item on the instrument.

A. The Value of Risk Communication in Justice Settings

The justice system has not embraced the principles of risk communication that are used in medicine and other fields to facilitate informed decisions by the justice-involved individual. For example, consumer credit scores are often accompanied by a communications strategy and analysis of how an individual's credit score may be improved. Risk communication can be summarized as follows (World Health Organization, 2021):

> Risk communication refers to the exchange of real-time information, advice and opinions between experts and people facing threats to their health, economic or social well-being. The ultimate purpose of risk communication is to enable people at risk to take informed decisions to protect themselves and their loved ones. Risk communication uses many communications techniques ranging from media and social media communications, mass communications and community engagement. It requires a sound understanding of people's perceptions, concerns and beliefs as well as their knowledge and practices. It also requires the early identification and management of rumors, misinformation and other challenges.

Risk communication is the process by which sensitive information is shared with individuals in a direct, nonjudgmental manner. This type of communication can raise awareness, encourage protective behavior, build a person's knowledge about hazards and risks, help the individual to accept risk factors and implement changes, guide an individual on how to address risk factors, and ensure that the individual understands that they are responsible for their own actions (Walters et al., 2014). Sharing difficult information can reduce uncertainty (which may create negative behavior) and improve the working relationship and trust between the individual and staff.

The field of health promotion and awareness has identified components of risk communication that are applicable in justice settings — since the justice setting acknowledges that deterrence is a goal, which implies that certain behaviors are desirable and normative and others are not. Risk communication practices include an emphasis on what information is communicated, how it is communicated, and who communicates the message. Messages are planned by the sender (the corrections agency) to convey what information the agency desires an individual to act on; that is, the messages can facilitate how behaviors, attitudes, or knowledge can be altered or influenced.

B. Theories of Risk Communication

The *communication persuasion* model, which is the classic model underpinning risk communication, looks at elements of the communication (who is the source, what is the message, what technique is used to convey the message, where is the message headed) and how it affects the steps to engage in attitudinal and/or behavior change (Glik, 2007). Studies have found that risk communication is more effective when: (1) the source of the information is perceived as credible, (2) the message is clear and concrete with clear outcomes, and (3) the message can resonate with the receiver. Quality messaging requires staff and the people on their caseloads to understand how the risk score is derived, and that the message around risk categories is vital to ensure that the messages are well-received. From a risk communication perspective, we have an obligation to give more emphasis and attention to methodological and

implementation factors that affect the quality of the instruments. Messages are more likely to have an impact when they are processed and understood, but the question is whether the sender is interested in the individual processing the message.

Crafting messages regarding information from RNA tools can also benefit from the health belief model. The more an individual perceives the risks associated with their behaviors and attitudes, the more likely the person is to act. Individuals are more likely to take proactive action steps when they believe that the risks are real and that their actions will result in more positive behaviors. This model also acknowledges that the benefits of behavior change must be outweighed by the costs (barriers). Legitimacy of justice actions (Sunshine and Tyler, 2003) and decisions can thus affect the benefit-cost calculations — individuals who perceive that their behavior change will not be acknowledged or respected by the justice system. This is compounded by the prevailing concern that the justice system is more interested in "lock 'em up" practices than behavioral change.

Protection motivation theory offers the calculation that threat and coping appraisals are important in shaping a person's desire to protect themself (Glik, 2007). In regard to the justice system, this means that individuals are determining that the intent of the justice system is positive and that being prepared (through one's behavior) can mitigate a coping response. Desired outcomes can be learned through mimicking others, particularly if the person has the skills, has self-efficacy, and assesses that the outcome will be beneficial.

Stages of change (precontemplation, contemplation, action, maintenance, etc.) (Prochaska, DiClemente, and Norcross, 1992) can be used to assess where the individual is in regard to their expectations of outcomes. The stage of change model not only suggests that different messages are desirable in each stage, but it also suggests that the type of action taken by an individual will vary. The justice system needs to recognize the variations in types of responses by the stage, and consider these responses to be legitimate. Thus, an individual in the precontemplation stage may not be ready to accept their risk level but may be ready to learn more about what factors affect that risk level, whereas an individual in the action phase may be working to address risk factors. It is critical to acknowledge that individuals pass through various stages, which directly affects their responses — from engagement to knowledge seeking to compliance to aggressive actions. Communicating the acceptance for an individual's actions, while recognizing this stage, is important to improve the credibility of the source.

Targeting messages to different audiences promotes utilization of the message. Messages that are specific to each audience should be designed, tested, and used. Responsivity factors (gender, mental health needs, racial and cultural makeup, etc.) might be considered for different types of audiences. We know that younger individuals have different responses to certain messages than others given psychological and physical maturation issues. This suggests a need to discuss risk categories differently to increase the credibility of the information.

Therefore, the risk communication literature explains that merely providing a person with a risk score is insufficient to help them understand the meaning of the score, how to use it to change an attitude or behavior, or help them learn how to make such a change. Given that there is a power differential between staff and the individual, a shared decision-making model is needed where individuals have a voice and their voice matters in making choices for their case plans. This is consistent with persuasion approaches (Matejkowski, Lee, and Severson, 2018).

The model for shared decision-making requires officers to share RNA information and then engage the individual in a discussion about a supervision/case/treatment plan where the individual has the authority to make decisions.

C. Transmitting Information

Transmitting information is key to risk communication. Increasingly, information that visualizes the problem can be shared instead of using words to describe a problem. Appendix A contains a series of figures that depict different ways in which information can be visualized. The surge in web-based therapies, cell phones, and apps increases enthusiasm for visual messages and multimedia messages. Using graphics or techniques to convey what the risk score means would greatly enhance communication. In a recent study that employed a web-based intervention to share motivational messages (Walters et al., 2014), a variety of graphics were used to convey complicated information to encourage people to understand their own risk behaviors. The study found that the visual messages had an impact on individuals initiating treatment over standard intake processes where information is collected through an RNA (Lerch et al., 2017). The web-based tool improved engagement in treatment, identified key factors that motivated individuals to change, and was cost-effective (Lerch et al., 2017; Cowell et al., 2018; Spohr et al., 2017; see https://youtu.be/58nDBwlHmvY).

D. Comprehensive Communication System

A total risk communication system integrates information from the RNA and links it to part of the process by which correctional control exists. That is, besides relying upon staff to transmit information from the RNA, the goal should be to build the RNA information into various steps of the process. A good communication system helps guide an individual to complete the instrument and can: (1) tailor feedback to the offender; (2) provide reflections, information, and suggestions; and (3) link individual responses to help a person make decisions on various choices that are consistent with being under correctional control. It can provide feedback loops to help individuals link their responses to certain risk and need factors. This reduces the burden on the individual to integrate and evaluate numerous predictors, and it allows the individual to think about goals and objectives for themselves based on their risk to recidivate and their own criminogenic needs.

The advantage of a risk communication system, as compared to merely an instrument, is that it provides an integrated model to support decision-making. The following examples illustrate how information can be transmitted and integrated into different types of supervision or correctional control processes. Currently, we are not aware of any U.S. correctional agencies that use a risk communication system; however, such a system would be useful to advance the implementation and use of RNA — an innovation that is needed in the field. Risk communication focuses attention on the messages conveyed to facilitate behavioral change; current practice is focused more on using a tool. Examples of what can be brought into the risk communication system might include:

- **Case plan improvements.** Taxman and Caudy (2015) illustrate some typologies of patterns of offenders, which are then suitable for correctional plans, supervision protocols, or practice guidelines. Such typologies can be generated from the RNA system using various statistical methods. Algorithms can be used to define typologies and to help create prototype supervision or

Table 5: Example of Possible Typologies

Typology	Definition	Treatment	Controls	Incentives	Sanctions
Violent	Personal crimes/ antisocial patterns High to moderate risk Four criminal needs, three destabilizers	Urban – CBT Rural – Telehealth, CBT	House arrest Electronic monitors	Participation in treatment for three months, reduce supervision by one month	Short-term jail Graduated sanctions
Intimate Partner Violence (IPV)	Pattern of IPV Personal crimes All risk levels, destabilizers, family issues	Urban – MST/CBT Rural – Social worker Family therapy for children	Restraining orders if needed Rural – Report in with sheriff	Participation in treatment for three months, reduce supervision by one month	Short-term jail Graduated sanctions
Opioid Substance Abusers	Diagnosed opioid user All risk levels, substance abuse disorders, housing issues	Urban – MAT CBT/telehealth Drug court Rural – Drug court	Drug testing PO visits	MAT retention for three months, reduce supervision by one month	Status hearings with court Short-term jail Residential treatment

Notes: The definition should be data-driven based on risk-need profile and confluence of need factors. CBT = cognitive-behavioral therapy, MAT = medication-assisted treatment, MST = multisystemic therapy, PO = probation officers.

case management plans, as shown in Table 5.

- **Treatment-matching algorithms.** The RNA information and typologies can be supported by algorithms that match the individual to the appropriate programs based on the program's key characteristics. This will allow individuals to view the category(ies) of treatment that are most appropriate. The list of computer-generated treatment matches may help individuals engage in those services (Taxman and Pattavina, 2013).

- **Case-planning component.** Case management can help individuals outline their short- and long-term goals to reduce recidivism with specific actions steps. The case-planning tool can begin with requirements, such as drug testing and frequency of face-to-face contacts, and then focus on attending to risk and need factors.

- **Providing for continuous planning process.** A continuous process of case planning has the individual work between treatment sessions geared toward their goals of reducing recidivism. This builds in an extender model where a person completes progress reports before their next appointment, and these progress reports are the discussion points for the supervision meeting.

- **Graduated sanctions/incentives components.** Based on studies that demonstrate that incentives are more effective than sanctions in eliciting behavioral change (Sloas et al., 2019; Mowen, Wodahl, and Garland, 2018), the integrated process should build in alerts that acknowledge progress and gains. In addition, the system can highlight when individuals are not complying.

4. RNA tools can achieve fairness-related goals and the greater good by using risk communication strategies that link the RNA information to action steps by the individual.

    - Training the correctional staff who will be using the RNA tool is essential for effective communication, particularly in how to explain the needs and translate it into a case plan.

    - A risk communication system provides an integrated model for decision-making.

        — This system includes supervision plan improvements, treatment-matching algorithms, and graduated sanctions and incentives.

        — An integrated model helps increase the individual's awareness of their own circumstance and need for programming.

The principles and guidelines identified in this paper represent a relatively innovative approach to the design and use of RNA instruments in corrections. We are not aware of any jurisdiction in the U.S. that has fully applied all of the guidelines included within the four principles. If these guidelines were fully implemented, however, we anticipate that RNA instruments would be more reliable, efficient, and effective, particularly with regard to improved predictive performance. We believe that full implementation would lead not only to more responsible and ethical use of RNA tools but also to better, more equitable outcomes for correctional populations and systems.

Although our paper has focused primarily on risk assessment, we believe a similar review is necessary for needs assessment. Recently, a number of concerns have been raised about the domains of the needs assessment categories in RNA tools due to poor preprocessing and in-processing methods (see Ward and Carter, 2019; Ward and Fortune, 2016; Taxman and Smith, In press). If needs assessment received the same level of scrutiny that has recently been applied to risk assessment, we anticipate it would lead to further improvements in the design and implementation of RNA tools and improve decisions that are made.

To help facilitate use of the principles and guidelines we have identified in this paper, the following checklist illustrates the main points for development and utilization. Among the areas to consider, there are different responsibilities for the developers who design and validate the RNA instruments and the practitioners who use them.

RNA Guidelines That Could Be a Basis for Industry-Based Guidelines

RNA Development (Fairness, Efficiency, Effectiveness, Communication)

1. Identify how the target population, decision point (post-sentencing, intake, reassessment, etc.), and setting (jail, prison, probation, parole, etc.) may affect the design and/or use of the instrument.	☐
2. Consider the development of a customized RNA tool based on local data (instead of an off-the-shelf tool).	☐
3. For developers and practitioners, invest in the use of automated scoring processes for the RNA instrument to improve the instrument's efficiency.	☐
a. Design the RNA tool by including items that can be scored through an automated process.	☐
b. If the tool is designed to be scored through an automated process, identify the items that come from source data and identify any proxy variables.	☐
c. If variables are proxies for other variables, identify the variables and demonstrate that the proxy is appropriate. Include as much data as feasible.	☐
4. Encourage the use of standardized needs assessment tools for dynamic need factors.	☐
a. If possible, use more efficient processes, such as computer-assisted surveys in which justice-involved people complete their own needs assessments.	☐
b. If computer-assisted survey technology is used, it should be piloted on the correctional population for which it will be used.	☐

RNA Validation (Fairness, Efficiency, Effectiveness, Communication)

1. Examine the full range of available algorithms (including machine learning algorithms) used to develop the instrument, and document the algorithms used.	☐
2. If the RNA instrument must be manually scored by correctional staff, conduct an inter-rater reliability study prior to using the instrument.	☐
a. *For practitioners:* Coach staff on any items that are scored inconsistently, as shown by the inter-rater reliability assessment.	☐
b. *For practitioners:* Develop a routine annual inter-rater reliability assessment to identify proper use of the instrument.	☐
3. Revalidation should happen at least once every 5 years, or when the instrument is used on a new/different population.	☐
4. *For practitioners:* Discuss the implications of these procedures on the results, particularly on creating risk categories or classification.	☐

RNA Implementation (Fairness, Efficiency, Effectiveness, Communication)

1. *For practitioners:* Create graphics to illustrate risk information for clients to help them understand static and dynamic risk factors.	☐
2. *For practitioners:* Develop policies to make sure staff are following the risk principle in making program assignment decisions.	☐
3. *For practitioners:* Demonstrate how the RNA instrument should be used as part of correctional and/or supervision decisions to improve fair and just decision-making models.	☐
4. *For practitioners:* Ensure staff are provided with communication strategies to emphasize sharing information.	☐
a. Communication strategies should focus on shared decision-making, making joint decisions, and dealing with competing risk and need factors.	☐
b. Ensure staff use communication strategies that emphasize encouragement and motivation to tackle risk and need behaviors.	☐
5. Provide regular boosters or reinforcers for staff to ensure they are comfortable with motivational language and how to handle difficult situations.	☐

Documentation To Facilitate Transparency

Development

1. Map data sources used in the RNA tool, identifying how different criminal justice events (arrest, type of arrest, conviction, etc.) are handled and how each variable will be measured.	☐
a. *For practitioners:* Share the map of the data sources with users to help them understand what data are included in the RNA tool and how they were measured.	☐
b. *For practitioners:* Review the data map and identify variables that could be adjusted to improve the tool's validity.	☐
2. Prepare a Data Definition Guide that describes how each item on the RNA instrument is measured.	☐

Validation

1. Prepare a report on the development and validation of the RNA instrument.	☐
a. Report predictive performance statistics from the test set (or validation sample).	☐
b. Calculate and present key statistics for each risk model considered, including false positive and false negative rate, sensitivity, and specificity.	☐
c. Assess predictive performance across groups (gender, race, ethnicity, etc.) and various algorithms considered to promote fairness and equity.	☐
• While some differences among groups may be expected, researchers and practitioners should assess how these differences affect the tool's validity, both in terms of face validity and predictive validity.	☐
• Document whether adjustments were made to address group differences, and describe how any adjustments affect the tool's validity.	☐
2. Document both the training and validation datasets on the procedures that were used to address any differences that may occur among the groups (gender, race, etc.).	☐
3. *For practitioners:* Share RNA documentation widely to increase knowledge in the instrument and trust about the validity of the design process.	☐
4. *For practitioners:* Examine the researcher's documentation of the training and validation datasets to ensure that there is agreement on how the differences were addressed.	☐

Training Materials	
1. *For researchers and practitioners:* Develop training material to help staff understand the meaning behind each component of the RNA.	☐
2. Provide training and booster sessions.	☐

Protocols	
1. Establish agency protocols for sharing information with other agencies and with individuals who are under justice control.	☐
2. Establish agency protocols for RNA utilization.	☐

References

Andrews, D.A. and Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law* 16: 39-55.

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

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (May 23, 2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals, And It's Biased Against Blacks. ProPublica.

Austin, J., Coleman, D., Peyton, J., and Johnson, K. (2003). Reliability and Validity Study of the LSI-R Risk Assessment Instrument. Unpublished study submitted to the Pennsylvania Board of Probation and Parole.

Benson, B. (2016). Cognitive Bias Cheat Sheet. https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18.

Beretta, E., Santangelo, A., Lepri, B., Vetro, A., and DeMartin, J.C. (2019). *The Invisible Power of Fairness. How Machine Learning Shapes Democracy.* Turin, Italy: Nexa Center for Internet & Society.

Berk, R. (2019). Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies* 16(1): 175-194.

Berk, R. and Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy* 12(3): 513-544.

Berk, R. and Elzarka, A. (2020). Almost politically acceptable criminal justice risk assessment. *Criminology & Public Policy* 19(4): 1231-1257.

Berk, R., Hidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 1-42. doi: 10.1177/0049124118782533.

Blasko, B.L., Souza, K.A., Via, B., DelPrinciple, S., and Taxman, F.S. (2016). Performance measures in community corrections: Measuring supervision practices with existing agency data. *Federal Probation* 80(3): 26-31. http://www.uscourts.gov/sites/default/files/usct10024-fedprobation-dec2016_0.pdf.

Bonta, J. and Andrews, D.A. (2007). *Risk-Needs-Responsivity Model for Offender Assessment and Rehabilitation*. Ottawa, Ontario, Canada: Public Safety Canada.

Breiman, L. (1996). Bagging predictors. *Machine Learning* 26: 123-140.

Breiman, L. (2001). Random forests. *Machine Learning* 45: 5-32.

Breitenbach, M., Dieterich, W., Brennan, T., and Fan, A. (2009). "Creating Risk-Scores in Very Imbalanced Datasets: Predicting Extremely Low Violent Crime Among Criminal Offenders Following Release from Prison," in *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection,* ed. Y.S. Koh and N. Rountree. Hershey, PA: Information Science Reference.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36: 21-40.

Bucklen, K.B. (2007). The Predictive Validity of Two Risk Assessment Tools Among a Sample of Pennsylvania Offenders. Unpublished study prepared for the Pennsylvania Department of Corrections.

Burgess, E.W. (1928). "Factors Determining Success or Failure on Parole," in *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois,* ed. A.A. Bruce, E.W. Burgess, J. Landesco, and A.J. Harno. Springfield, IL: Illinois State Board of Parole, 221-234.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble Selection From Libraries of Models. Paper presented at the proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics* 4(1): 266-298.

Corbett-Davies, S., Pierson, E., Feller, A., Goael, S., and Huq, A. (2017). Algorithmic Decision Making and Cost of Fairness. KDD 2017 Research Paper. Halifax, Nova Scotia, Canada. doi: 10.1145/3097983.3098095.

Cowell, A., Zarkin, G., Wedehase, B.J., Lerch, J.A., Walters, S., and Taxman, F.S. (2018). Cost and cost-effectiveness of computerized vs. in-person motivational interventions in the criminal justice system. *Journal of Substance Abuse Treatment* 87(2): 42-49.

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

Doleac, J. and Stevenson, M. August 22, 2016. "Are Criminal Justice Risk Assessment Scores Racist?" Brookings Institute. https://www.brookings.edu/blog/up-front/2016/08/22/are-criminal-risk-assessment-scores-racist/.

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

Duwe, G. (2014). The development, validity, and reliability of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR). *Criminal Justice Policy Review* 25: 579-613.

Duwe, G. (2017). Better practices in the development and validation of recidivism risk assessments: The Minnesota Sex Offender Screening Tool-4. *Criminal Justice Policy Review.* https://doi.org/10.1177%2F0887403417718608.

Duwe, G. and Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Review* 28(6). doi: 10.1177/0887403415604899.

Duwe, G. and Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections: Policy, Practice and Research* 1: 155-176.

Duwe, G. and Rocque, M. (2017). The effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy* 16: 235-269.

Duwe, G. and Rocque, M. (2018). The home-field advantage and the perils of professional judgment: Evaluating the performance of the Static-99R and the MnSOST-3 in predicting sexual recidivism. *Law and Human Behavior* 42(3): 269-279.

Duwe, G. and Rocque, M. (2019). The predictive performance of risk assessment in real life: An external validation of the MnSTARR. *Corrections: Policy, Practice and Research.* doi.org/10.1080/23774657.2019.1682952.

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38: 367-378.

Glik, D.C. (2007). Risk communication for public health emergencies. *Annual Review of Public Health* 28(1): 33-54.

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

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8: 23-34.

Hamilton, Z., Neuilly, M.-A., Lee, S., and Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology* 11(2): 299-318.

Hand, D.J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning* 77: 103-123.

Hess, J. and Turner, S. (2013). *Risk Assessment Accuracy in Corrections Population Management: Testing the Promise of Tree Based Ensemble Predictions.* Irvine, CA: The University of California, Center for Evidence-Based Corrections.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

Lerch, J.L., Tang, L., Walters, S., and Taxman, F.S. (2017). Effectiveness of a computerized motivational intervention on treatment initiation and substance use: Results from a randomized trial. *Journal of Substance Abuse Treatment* 80: 59-66.

Lin, Z., Jung, J., Goel, S., and Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances* 6(7).

Liu, Y.Y., Yang, M., Ramsey, M., Li, X., and Coid, J.W. (2011). A comparison of logistic regression, classification and regression tree, and neural network models in predicting violent re-offending. *Journal of Quantitative Criminology* 27: 547-573.

Magnuson, S., Kras, K., Aleandro, H., Rudes, D., and Taxman, F.S. (2019). Using plan-do-study-act and participatory action research to improve use of risk needs assessments. *Corrections: Policy, Practice and Research* 5(1): 44-63. doi: 10.1080/23774657.2018.1555442.

Martinson, R. (1974). What works? Questions and answers about prison reform. *The Public Interest* 34: 22-54.

Matejkowski, J., Lee, S., and Severson, M. (2018). Validation of a tool to measure attitudes among community corrections officers toward shared decision making with formerly incarcerated persons with mental illness. *Criminal Justice and Behavior* 45: 612-627.

Mayson, S.G. (2019). Bias in, bias out. *The Yale Law Journal* 128: 2218-2300.

Meehl, P.E. (1954). *Clinical Versus Statistical Prediction.* Minneapolis, MN: University of Minnesota Press.

Miller, J. and Maloney, C. (2013). Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. *Criminal Justice and Behavior* 40: 716-736.

Mowen, T.J., Wodahl, E., and Garland, B. (2018). The role of sanctions and incentives in promoting successful reentry: Evidence from the SVORI data. *Criminal Justice and Behavior* 45(8): 1288-1307.

Pretrial Justice Institute. (2020). *Updated Position on Pretrial Risk Assessment Tools.* Gaithersburg, MD: Pretrial Justice Institute. https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf.

Prochaska, J.O., DiClemente, C.C., and Norcross, J.C. (1992). In search of how people change: Applications to addictive behaviors. *American Psychologist* 47(9): 1102-1114.

Rice, M.E. and Harris, G.T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior* 29(5): 615-620.

Ridgeway, G. (2013). "The Pitfalls of Prediction." *NIJ Journal* 271, February 2013.

Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analyses. *The Knowledge Engineering Review* 29(5): 1-54. doi: https://doi.org/10.1017/S0269888913000039.

Simourd, D. (2006). Validation of Risk/Needs Assessments in the Pennsylvania Department of Corrections. Unpublished study prepared for the Pennsylvania Department of Corrections.

Sloas, L., Wooditch, A., Murphy, A., and Taxman, F.S. (2019). Assessing the use and impact of points and rewards across four federal probation districts: A contingency management approach. *Victims & Offenders* 14(7): 811-831.

Spohr, S., Walters, S., and Taxman, F.S. (2017). People's reasons for wanting to complete probation: Use and predictive validity in an e-health intervention. *Evaluation & Program Planning* 61: 144-149.

Sunshine, J. and Tyler, T.R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. *Law & Society Review* 37: 513-548. doi: 10.1111/1540-5893.3703002.

Taxman, F.S. (2017), ed. *Handbook on Risk and Need Assessment: Theory and Practice*. Milton, England: Routledge Press.

Taxman, F.S. (2018a). The partially clothed emperor: Evidence-based practices. *Journal of Contemporary Criminal Justice* 34(1): 97-114.

Taxman, F.S. (2018b). "Risk and Needs Assessment: Moving Forward with Improved Methods," in *Handbook on Risk Assessment,* ed. J. Singh. London: Oxford University Press.

Taxman, F.S. and Ainsworth, S. (2009). Correctional milieu: The key to quality outcomes. *Victims & Offenders* 4(4): 334-340.

Taxman, F.S. and Caudy, M. (2015). Risk tells us who, but not what or how: Empirical assessment of the complexity of criminogenic needs to inform correctional programming. *Criminology & Public Policy* 14(1): 71-103.

Taxman, F.S. and Pattavina, A. (2013). *Simulation Strategies to Reduce Recidivism: Risk Need Responsivity (RNR) Modeling in the Criminal Justice System*. New York: Springer.

Taxman, F.S. and Smith, L. (In press). Risk-Need-Responsivity (RNR) classification models: Still evolving. *Aggression and Violent Behavior*. https://www.sciencedirect.com/science/article/abs/pii/S1359178920301634.

Taxman, F.S., Walters, S.W., Sloas, L., Lerch, J., and Rodriguez, M. (2015). Counselor vs. computer: A randomized controlled trial protocol comparing motivational tools to improve probationer treatment initiation and reduce substance use. *Contemporary Clinical Trials* 43: 120-128.

Thurman, T., Chowdhury, S., and Taxman, F.S. (2019). Fidelity measures for risk-need assessment (RNA) tools usage in case plans. *Corrections: Policy, Practice and Research*. doi: 10.1080/23774657.2019.1696252.

Tollenaar, N. and Van der Heijden, P. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning, and data mining predictive methods. *Journal of the Royal Statistical Society,* Series A 176 (part 2): 565-584.

Toronjo, H. and Taxman, F.S. (2017). "Supervision Face-to-Face Contacts: The Emergence of an Intervention," in *Evidence-Based Skills in Community Justice: International Perspectives on Effective Practice,* ed. P. Ugwudike, J. Annison, and P. Raynor. Bristol, England: The Policy Press.

U.S. Department of Health and Human Services, National Institutes of Health, National Institute on Aging. (2011). Harmonization Strategies for Behavioral, Social Science, and Genetic Research: Workshop Summary. Bethesda, MD: U.S. Department of Health and Human Services. https://www.nia.nih.gov/sites/default/files/d7/nia_bssg_harmonization_summary_version_2-5-20122.pdf.

U.S. Department of Justice, Office of the Attorney General. (2019). *First Step Act of 2018: Risk and Needs Assessment*. Washington, DC: U.S. Department of Justice, Office of the Attorney General.

Vapnick, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Viglione, J. (2019). The risk-need-responsivity model: How do probation officers implement the principles of effective intervention? *Criminal Justice and Behavior* 46(5): 655-673.

Viglione, J., Rudes, D.S., and Taxman, F.S. (2015). Misalignment in supervision: Implementing risk/needs assessment instruments in probation. *Criminal Justice and Behavior* 42(3): 263-285.

Viglione, J. and Taxman, F.S. (2018). Low risk offenders under probation supervision: Risk management and the risk-needs-responsivity (RNR) framework. *Criminal Justice and Behavior* 45(12): 1809-1831. https://doi.org/10.1177/0093854818790299.

Walters, S.W., Ondersma, S.J., Ingersoll, K.S., Rodriguez, M., Lerch, J., and Taxman, F.S. (2014). MAPIT: Development of a web-based intervention targeting substance abuse treatment in the criminal justice system. *Journal of Substance Abuse Treatment* 46(1): 60-65.

Ward, T. and Carter, E. (2019). The classification of offending and crime related problems: A functional perspective. *Psychology, Crime & Law* 25(6): 542-560. https://doi.org/10.1080/1068316X.2018.1557182.

Ward, T. and Fortune, C.-A. (2016). The role of dynamic risk factors in the explanation of offending. *Aggression and Violent Behavior* 29: 79-88. https://doi.org/10.1016/j.avb.2016.06.007.

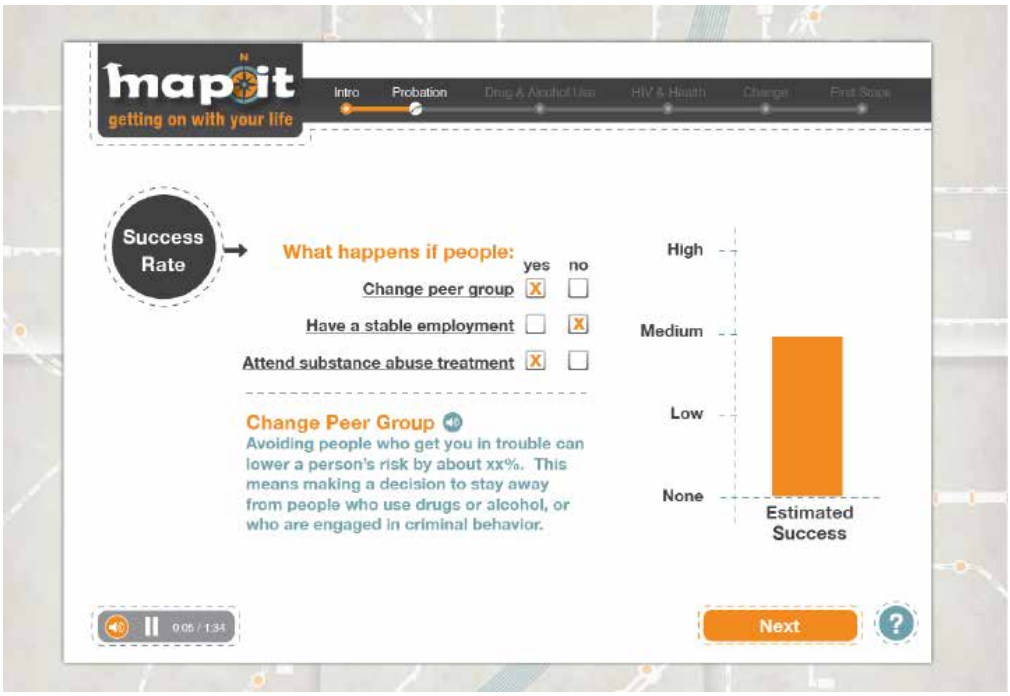
World Health Organization. (2021). Risk Communication. https://www.who.int/risk-communication/background/en/.

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

Helping Individuals Understand Their Areas of Concern

This graphic helps an individual understand the factors that are affecting their risk score.

Helping Communicate How Changes Can Affect Risk and Success

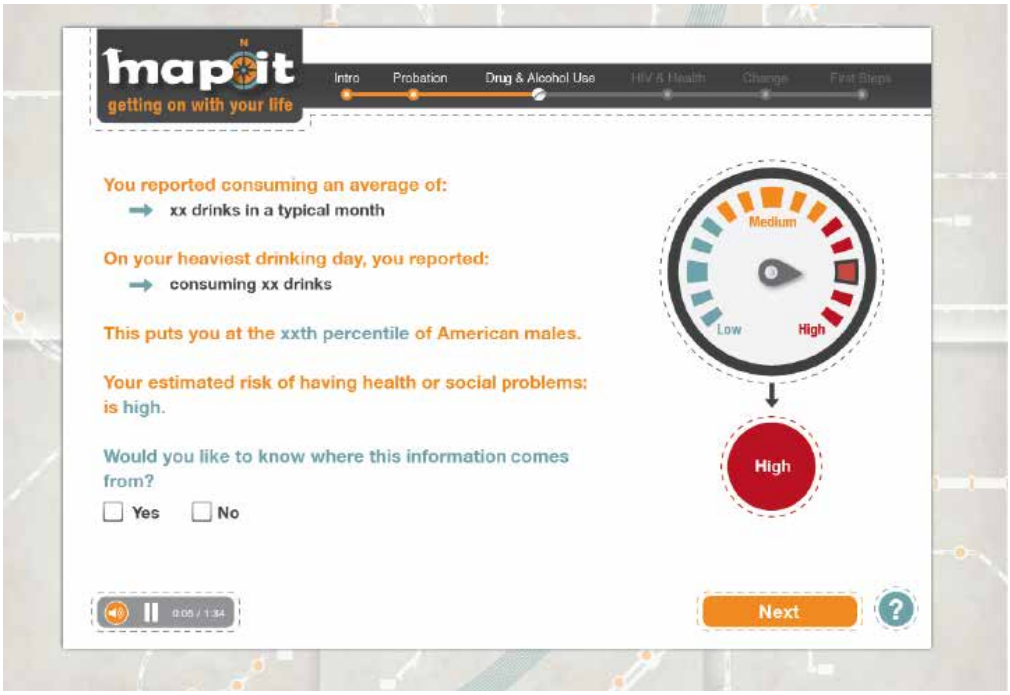


This graphic helps individuals assess the factors they can address that will improve their chances for success.

Helping Individuals Compare Their Behavior to Others in the Same Gender and Age Groups

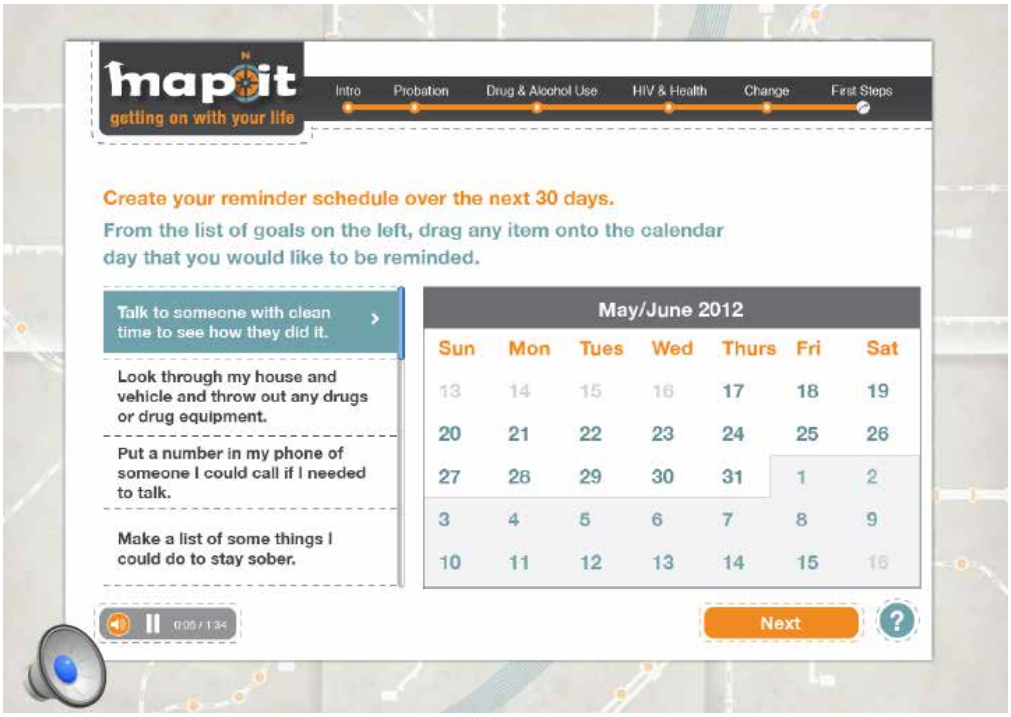
This graphic helps an individual review their own behaviors in comparison to other individuals in a similar age and gender category. It helps an individual assess their behavior overall, not only as part of their social network.

Communicating the Severity of the Behavior



This graphic indicates the degree to which the behavior is problematic. It shows an individual that certain behaviors are more severe than others.

Setting Goals and Timelines



This graphic can guide individuals on how to set up goals and target behaviors to address the goals.