Challenge Accepted? A Critique of the 2021 National Institute of Justice Recidivism Forecasting Challenge

Tobi Jegede
American Civil Liberties Union
Washington, DC, USA
tjegede@aclu.org

Marissa Kumar Gerchick
American Civil Liberties Union
New York City, NY, USA
mgerchick@aclu.org

Amreeta S. Mathai
New York Civil Liberties Union
New York City, NY, USA
amathai@nyclu.org

Aaron Horowitz
American Civil Liberties Union
San Francisco, CA, USA
ahorowitz@aclu.org

ABSTRACT

In 2021, the National Institute of Justice — the research arm of the United States Department of Justice — released the "Recidivism Forecasting Challenge" ("the Challenge") with the stated goals of "increas[ing] public safety and improv[ing] the fair administration of justice across the United States," providing "critical information to community corrections departments...," and ultimately "improv[ing]" the ability to forecast recidivism using person-and place-based variables" [68]. The Challenge was also designed, in part, to encourage "non-criminal justice forecasting researchers to compete against more 'traditional' criminal justice researchers" [68]. Challenge contestants had the opportunity to win part of the $723,000 in prize money for their submitted models. In this work, we highlight how the Challenge was underpinned by a techno-solutionist framing (emphasizing technical interventions without addressing underlying structural problems) [66] and plagued by serious ethical and methodological issues, including (1) the choice of training data and the selection of an outcome variable extracted from racially biased and inaccurate law enforcement data systems, (2) data leakage that may have seriously compromised the Challenge, (3) the choice of a faulty fairness metric, leading to the inability of submitted models to accurately surface any bias issues in the data selected for the Challenge, (4) the inclusion of candidate variables that created the potential for feedback loops, (5) a Challenge structure that arguably incentivized exploiting the metrics used to judge entrants, leading to the development of trivial solutions that could not realistically work in practice, and (6) the participation of Challenge contestants who demonstrated a lack of understanding of basic aspects of the U.S. criminal legal system's structure and functions. We analyze the Challenge and its shortcomings through the lens of participatory design, applying emerging principles for robust participatory design practices in artificial intelligence (AI) and machine learning (ML) development to evaluate the Challenge's structure and results. We argue that if the Challenge's designers had adhered to these principles, the Challenge would have looked dramatically different or would not have occurred at all. We highlight several urgent needs and potential paths forward for any future efforts of this nature, recognizing the real and significant harms of recidivism prediction tools and the need to center communities directly impacted by policing and incarceration when thinking about whether to develop risk assessment tools.

CCS CONCEPTS

• **Human-centered computing** → *Interaction design theory, concepts and paradigms*; • **Applied computing** → **Law**.

KEYWORDS

participatory design, crowdsourcing, recidivism, criminal justice, algorithmic design, risk assessment

ACM Reference Format:
Tobi Jegede, Marissa Kumar Gerchick, Amreeta S. Mathai, and Aaron Horowitz. 2023. Challenge Accepted? A Critique of the 2021 National Institute of Justice Recidivism Forecasting Challenge. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3617694.3623242

1 INTRODUCTION

In April 2021, the National Institute of Justice ("NIJ") released the Recidivism Forecasting Challenge ("the Challenge") that sought to "increase public safety and the fair administration of justice by improving the ability to forecast and understand the variables that impact the likelihood that an individual under parole supervision will recidivate" [68]. The Challenge organizers designed the Challenge with the aim to encourage "non-criminal justice forecasting researchers to compete against more 'traditional' criminal justice researchers" for a chance to win part of the $723,000 in cash prizes [68]. At the conclusion of the Challenge, challenge winners were asked to write papers describing the process of developing their winning models and to provide commentary on the Challenge itself. While heralded by the NIJ as a successful effort that "demonstrate[d] the value of open data and open competition" [50], in reality, the Challenge was marked by serious and fundamental flaws. One of the winning papers came to the following conclusion: "We are hesitant to accept any insights gained from submitted models and question the reliability of their performance. We would also discourage the

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
EAAMO '23, October 30–November 01, 2023, Boston, MA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0381-2/23/10.
https://doi.org/10.1145/3617694.3623242

use of any submitted models in live environments" [18]. This paper was not an anomaly — 6 of the other 25 winning papers also questioned the reliability and practical utility of the Challenge's results [22, 56, 67, 92, 98, 100].

In this work, we highlight shortcomings with the Challenge that can be viewed as failures to implement robust participatory design practices — practices that center those who would be impacted by or would interact with an algorithm in the development of the algorithm [83], including in the evaluation and selection of training data. Although not explicitly advertised as a challenge focused on principles of participatory design, several characteristics of participatory design thinking were present in the framing of the Challenge, including the NIJ's solicitation of participation from non-criminal legal system[1] professionals in an effort to increase the "diversity of expertise and individuals evaluating the data" [50]. Despite calling for a diversity of experts, the kind of participatory design that the Challenge organizers engaged in could be construed as a form of participation-washing [83, 84], where the "diverse" group of individuals brought in to design an algorithmic tool are not the ones who would be affected by the tool, lack the context-specific knowledge to understand the limitations and problems with the proffered output and input variables, and may not fully understand what deployment of an algorithmic tool in the particular domain would look like. We assert that a lack of robust participatory design practices, like a flawed choice of challenge participants and a lack of meaningful inclusion of impacted communities, amplified methodological issues with the Challenge. We find that if tools built with poor participatory design practices are put into real-life environments, they could produce real and significant harms for the communities onto which they are deployed, as risk assessment tools currently used in the criminal legal system already do today [55]. In the following sections, we (1) situate the NIJ's Recidivism Forecasting Challenge in the broader ecosystem of predictive analytics tools used in the criminal legal system, (2) provide a set of guidelines for assessing what robust participatory design could look like and evaluate the Challenge against these guidelines, and (3) conclude by highlighting other methodological issues with the Challenge and discussing the policy implications of the Challenge.

2 RELATED WORK

2.1 Predictive Analytics Tools In the Criminal Legal System

The Challenge fits squarely into an ever-growing body of work that focuses on the increasing use of several types of predictive analytics tools in the criminal legal system [20, 29, 81, 93, 99], including risk assessment tools. Risk assessment tools in the criminal legal system have been used to make decisions about whether people should be subject to pre-trial detention [24, 52, 88], to determine the length of peoples' incarceratory sentences [65, 87], and to predict a person's likelihood to recidivate [21, 29]. The Challenge fits into the existing literature about the use of risk assessment tools to predict recidivism. The NIJ asked contestants to develop a risk assessment tool that would estimate the probability that individuals released from incarceration under parole supervision would be arrested for a new felony or misdemeanor crime within three years of the start of their release [68].

2.1.1 Critiques of Recidivism Risk Assessment Tools. The increased use of recidivism risk assessment tools in the criminal legal system has been met with criticism. While not an exhaustive list, some of the critiques highlighted in the literature include: (1) the fact these tools are often not equally accurate for people of different races [4, 21, 29, 93], (2) that the data used to build these tools is inaccurate and biased [39, 44, 62, 77, 79], (3) that faulty and insufficient proxies are chosen to measure recidivism [32, 36, 40, 57], (4) that the evaluation criteria for these models is insufficient [21, 23, 42, 60], and (5) that these types of models do not work well in practice and fail to achieve their goals of providing accurate measures of recidivism risk [31, 62].

With respect to the first critique, Desmarais et al. [29] find that despite historically making up less of the incarcerated population, white defendants are more likely to receive more accurate recidivism risk predictions from existing recidivism risk prediction tools than their Black and Brown counterparts. Related to the second critique, many studies have found that predictive risk assessment tools often re-encode existing biases [32, 46, 77]. Specifically, Richardson et al. [77] examined several predictive analytics tools built using data from police departments subject to consent decrees or federal monitoring and found that even when federal investigations identified corrupt, racially biased, or otherwise illegal policing practices, data reflecting these biases were rarely corrected in or removed from training datasets [77]. Consequently, the risk assessment tools developed using faulty data are more of a reflection of police misconduct and systematic bias of the policing system than an accurate measure of risk.

In reference to the third critique, recidivism risk assessment tools are broadly trying to predict the future commission of a crime. Two common metrics for measuring recidivism in the literature are re-arrest and re-conviction, which often means arrest or conviction for any type of crime, regardless of its severity. Though widely used as an outcome variable in risk assessment tools, and specifically used as the outcome metric in the Challenge, re-arrest is an inaccurate and biased proxy for commission of a crime [32, 36, 40, 57] because an arrest does not indicate that a crime was actually committed and is not subject to other independent evaluation [62]. Additionally, some uses of a re-arrest outcome variable are crime-severity agnostic and treat all arrests as equally risky behaviors. This means that accidentally walking onto private property, which could be charged as misdemeanor trespass, is treated as equally dangerous or risky as arrest for a violent physical assault. In the case of the Challenge, the outcome of interest that contestants were asked to optimize for was a crime severity agnostic re-arrest variable — arrest "for a new felony or misdemeanor crime within 3 years of the parole supervision start date" [68]. Using these kinds of outcomes runs the risk of building statistical models that reflect and may exacerbate historical biases [31] and may actually overestimate risk [88].

[1] Throughout this paper, we use the term "criminal legal system" rather than the criminal justice system to highlight that the system involves various methods of interaction with actors in the legal apparatus (police officers, judges, probation officers, etc.) and often does not provide justice for the communities who interact with the system. Rather, the system often disproportionately targets marginalized communities and creates lasting harms for those who interact with the system [9, 14, 80].

In reference to the fourth critique, the literature suggests that accuracy alone is not the best heuristic against which to measure the performance of a model [12]. Given the flawed and biased datasets that are often used to build recidivism risk assessment tools, fairness — the idea that a model should perform equally well across all sub-populations of interest (race, gender, etc.) — is offered up in the literature as a supplementary metric to accuracy to optimize for when building risk assessment tools. However, definitions of fairness are often incomplete. A fairness standard could allow for a faulty algorithm that performs equally poorly for all sub-populations of interest to still be found to be "fair" [23]. To move away from the difficulties of defining fairness mathematically, Green [41] advocates for "substantive algorithmic fairness" whereby algorithmic tool designers strive to "address upstream social disparities that feed into decision-making processes" and to "reduce downstream harms" without defining strict mathematical definitions of fairness that are not sufficiently translatable across different domains.

Finally, in reference to the fifth critique, when used in practice, many risk assessment tools do not work in the way that they are designed to function [46, 52, 62, 87]. Imai et al. [52] found that a pretrial risk assessment tool produced worse outcomes across race and gender, in terms of disparately measured levels of risk, than if the tool had not been used in the first place. Similarly, Stevenson and Doleac [87] find that a risk assessment tool they analyzed did not produce more accurate or fair results than human decision-makers, with younger defendants systematically receiving longer sentences than they would have received without algorithmic intervention.

In light of the aforementioned critiques, there are ongoing discussions in the literature about suggested remedies. Here, we simply highlight literature that provides thoughts about how one might improve recidivism risk assessment tools but we do not argue that these proposed solutions would be sufficient to deal with the multitude of issues plaguing recidivism risk assessment tools. Furthermore, we do not seek to legitimize the use of predictive analytics tools through the following discussion of proposed, theoretical remedies.

To address the issue of biased data, Gebru et al. [39] argue for the creation of information sheets about datasets that highlight how the data was collected and preprocessed, how the data should and shouldn't be used, and how the dataset will be stored. This documentation process, Gebru et al. [39] argue, will lead machine learning practitioners to be more accountable and more transparent about how the data developed for predictive analytics systems were generated. The hope through this process is that any latent issues with the dataset could be laid bare for the consumers of the dataset and people may choose not to use data with known biases and substantial flaws. Additionally, there is a category of suggested improvements to risk assessment tools that focuses on changing the design process for the development of such tools [3, 7, 47]. One of the suggested methods to improve the design of algorithmic tools is to crowdsource their development [3, 47]. Crowdsourcing has included using machine learning challenges or competitions that bring machine learning experts and non-experts into the healthcare [3], education [85], artificial intelligence [47], and criminal justice [68] fields to try to solve for a particular issue. An example of the use of crowdsourced challenges in the AI field is the use of bias bounties – competitions designed to surface embedded biases in deployed algorithmic tools [47]. The thinking, in this example and with other challenges more broadly, is that crowdsourcing the design of algorithmic tools produces better outcomes than an individual team would be able to produce and also allows for a variety of different models to be explored simultaneously [7]. In the following section, we expand on our discussion of crowdsourcing and talk about participatory design in machine learning contexts more broadly.

2.2 Participatory Design in Artificial Intelligence and Machine Learning

Discussions about the design, deployment, and evaluation of artificial intelligence and machine learning systems have increasingly focused on the idea of "participation" as a potential means to address some of the ethical and sociotechnical issues in AI development. The emphasis on participation is reflected in various arenas, from increasing scholarship examining and articulating participatory approaches to AI [11, 58, 84, 101], to federal guidance such as the White House's Blueprint for an AI Bill of Rights [48] ("the Blueprint") and the National AI Research Resource Task Force's goal to "address societal-level problems by strengthening and democratizing participation in...AI R&D" [38]. Here, we provide context around proposals for participatory approaches to the AI development process, identifying common principles and values for meaningful participation. This background grounds our evaluation, assessing whether the Challenge represented an opportunity for meaningful participation in line with these values.

Several works [11, 13, 84, 101] have highlighted that there is no agreed upon definition of what it means for an algorithmic design process to be "participatory." While the Challenge has not, to our knowledge, been explicitly described by the NIJ as an attempt at "participatory design," the NIJ's descriptions of its goals and functions invoke notions of participation. For example, the NIJ has described one of the functions of the Challenge as "help[ing] expand access to data and expertise" [50]. At the time of writing, the Challenge was also listed as an archived challenge on Challenge.gov,[2] the United States Government's "official hub for challenges and prize competitions across the U.S. federal government," which frames its aims with the language, "[l]earn how *you* can participate and make a difference" [3].

Researchers and practitioners have highlighted both potential strengths and weaknesses or risks of participatory design practices in the context of predictive tools [101]. Sloane et al. [84] and Birhane et al. [11] trace the history of participation through various mediums, including co-design practices in corporate settings in the 1970s, deliberative decision-making processes amongst formal and information organizations and groups across many time periods, and participation's current role in ML development processes. Birhane et al. [11] also trace the roots of participation to colonialist systems in the 1920s, highlighting how participation has been used to disguise harmful systems characterized by extreme concentrations of power under the guise of inclusivity. Sloane et al. [84] warn that some current forms of participation in machine learning design processes could amount to "participation washing" practices that

[2] See https://www.challenge.gov/?state=archived

are actually harmful to or extractive of the involved communities. Birhane et al. [11] similarly warn that participation can become a form of corporate "cooptation," or can improperly be presented as a panacea for broader issues around governance and inclusion. Chan et al. [19] highlight barriers to and problems with participation in AI development in the context of dataset development for AI and research labs working on AI. Yet these scholars also emphasize the opportunities that meaningful participation may hold for improving the status-quo [10, 11, 19, 84]. Several researchers have offered valuable principles for and frameworks of participatory design and offered myriad examples and case studies touching on the at times complicated and potentially beneficial aspects of participation [11, 28, 75, 78, 84, 86, 90].

In the realm of federal policy in the U.S., in 2022, the White House published the Blueprint, a guidance document outlining a set of principles and practices for the development of AI and ML systems. Among the Blueprint's principles is the tenet that people "should be protected from unsafe or ineffective systems," and the Blueprint emphasizes that consulting diverse stakeholders, including impacted communities, is central to realizing this principle. Further, this consultation must be paired with several robust protections including: (i) protections from "inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of" the reuse of such data [48, p. 15]. Additionally, the Blueprint highlights the need for norms regarding consent such that meaningful consent for data use is expected and respected and "[p]eople whose data is collected, used, shared, or stored by automated systems should be able to access data and metadata about themselves, know who has access to this data, and be able to correct it if necessary" [48, p. 35]. To institute these principles, the Blueprint discusses the importance of evaluating data for bias issues, taking into account the context around the data [48, p. 26], conducting proactive assessments to understand and address potential harms [48, p. 23], and more. Further, in 2023, the White House issued an Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government ("the Order") that directs federal agencies, including the Department of Justice, to address equity and systemic racism in U.S. policies and programs — including ending unjust disparities in the criminal legal system and "root[ing] out bias in the design and use of" artificial intelligence. In service of this goal, agencies are required to plan for "meaningful engage[ment] with underserved communities" and "incorporation of the perspectives of those with lived experiences into agency policies, programs, and activities" [49]. While the Blueprint and the Order were released after the Challenge concluded, they both serve as useful frameworks upon which to evaluate the Challenge. Furthermore, the Blueprint cites an extensive number of resources outlining and supporting its principles – including several resources released by the U.S. government prior to and during the Challenge.[3]

In this context, we use the next section of our paper to suggest a framework for what meaningful participatory design could look like in practice and examine the Challenge against this framework. We highlight the areas where the Challenge failed to be well-constructed and offer up insights on how future crowd-sourced machine learning challenges can learn from the mistakes of the Challenge.

[3]See the endnotes of [48] including endnotes 30, 62, and 78, which highlight examples of federal government resources related to algorithmic bias and data privacy released before the Challenge.

3 CRITIQUE OF THE NIJ CHALLENGE

3.1 Overview of the Challenge

To develop recidivism prediction models, the NIJ provided challenge contestants with a dataset that included 25,835 people who were released from Georgia prisons on discretionary parole to the custody of the Georgia Department of Community Supervision (GDCS) between January 1, 2013 and December 31, 2015. This dataset included demographic data, prior criminal history information, prison and parole case information, prior community supervision history, conditions of supervision as articulated by the Board of Pardons and Paroles, and supervision activities (violations, drug tests, program attendance, employment, residential moves, and accumulation of delinquency reports for violating conditions of parole) [68]. The models developed by contestants were judged based on two metrics meant to represent the accuracy and/or the fairness of their models. The accuracy metric the Challenge used was the Brier Score which is, in effect, a mean square error measurement specifically for determining the accuracy of probabilities, calculated using the following formula: $\frac{1}{n}\sum_{t=1}^{n}(f_t - A_t)^2$, where "n is the count of individuals in the test dataset, f_t is the forecasted probability of recidivism for individual t, and A_t is the actual outcome (0,1) for individual t" [68]. The fairness metric used was one minus the absolute value of the difference between the false positive rate for Black parolees and the false positive rate for white parolees: $FP = 1 - |FP_{Black} - FP_{White}|$. Contestants could chose to either just optimize their submitted models for accuracy or optimize for a metric that took both accuracy and fairness into account using a penalized fairness metric constructed in the following way: $(1 - BrierScore)(FP)$.

Challenge prizes were split across several categories. The prize categories were split up by outcome (predicting recidivism after 1 year on parole, after 2 years on parole or after 3 years on parole), sub-population (male vs. female parolees), and by evaluation metric (accuracy, as measured by the Brier Score, and fairness, as measured by a weighted fairness penalty). As an example, a contestant in the Challenge could chose to only submit a model that would predict recidivism for male parolees after 1 year on parole, optimizing for accuracy alone. If a contestant chose to just optimize for accuracy alone, then the goal was to achieve the lowest Brier score. Contestants who chose to optimize for fairness would win if their value for the penalized fairness metric described above was the highest among the other submitted models. At the end of the Challenge, the NIJ selected a total of 26 winners across the various outcome, sub-population, and evaluation metric categories [50].

3.2 Methods

In order to formulate a critique of the Challenge, we adapt frameworks and questions from Birhane et al. [11], Biderman and Scheirer [10], Sloane et al. [84], and the Biden-Harris Blueprint for an AI Bill of Rights [48] to synthesize a set of principles that we posit would be part of an effective participatory design process. We highlight

four main dimensions of participatory design for challenge designers to consider when building a crowdsourced machine learning challenge or tool including, (i) "Defining Participation", (ii) "Impacts of Participation", (iii) "Contesting Participation", and (iv) "Depth of Participation". We examine the extent to which the Challenge addressed these dimensions of participatory design. The framework presented below is not intended to be an exhaustive list of considerations for building robust crowd-sourced machine learning tools.

3.3 Elements of Meaningful Participatory Design

In this section, we analyze some of the tenets of participatory design in the context of the Challenge. We focus primarily on the experiences of (1) communities impacted by policing, arrest practices, and incarceration whose data may be used in the Challenge or who may be impacted by risk assessments deployed in the future and (2) contestants who entered the Challenge. Yet implicit in this Challenge is the participation of the many organizations and individuals involved with the use of risk assessments or the criminal legal system more broadly. This set extends beyond impacted communities and Challenge contestants to include law enforcement agencies (e.g., police, prosecutors, probation and parole officers, and NIJ itself, among others), public defenders, organizations offering assistance with probation and parole applications, other advocates for people impacted by the criminal legal system, and the ecosystem of researchers and corporations that profit or benefit from the extensive surveillance and incarceration in the criminal legal system, including those who design and sell risk assessment tools. While not the primary focus of this analysis, these people all crucially participate in the ecosystem upon which the Challenge is scaffolded.

3.3.1 Defining Participation. First, we focus on the question of: who is counted as a "participant" or decision-maker in the design process? (adapted from [10] and [11]). We conceive of three main types of participants in the Challenge – challenge organizers, contestants, and impacted communities. Challenge organizers include the NIJ and the law enforcement agencies/systems that collected, categorized, produced, and stored the data and that seek to deploy risk assessment tools in their work. Contestants are those who actively engaged in the Challenge and submitted models to the NIJ. Impacted communities include the people who are represented in the dataset itself. People who represent impacted communities such as public defenders and advocates for release on probation or parole could also be included in this broader category. We posit that people directly impacted by the design process, related data collection, and tool deployment should be decision-makers with actionable protections and meaningful say in the design process, including having an opportunity to comment on and contest the Challenge itself, to have their data removed from use in the Challenge, and to object to the use of the Challenge's results in future risk assessment development processes [75, 95].

At the core of the Challenge is the use of data about the experiences and lives of people released from prison in Georgia between 2013 and 2015 (the impacted community), and the potential for methods or models resulting from the Challenge to "provide critical information to community corrections departments" in the future (the organizers) [68]. From the Challenge's website and related documents, it is unclear if the people represented in this data were consulted in the design of the Challenge or were made aware of the Challenge at all. For a Challenge that revolves around data collection from communities impacted by incarceration and the results of which may be used in ways that affect those same communities, the meaningful participation (i.e., decision-making power and intentional consultation as opposed to participation washing [83, 84]) of impacted communities is essential. As highlighted by Petty et al. [75] in Our Data Bodies' 2018 report, which included interviews about data collection and digital privacy with more than 100 residents of historically marginalized communities in several U.S. cities, extractive data collection contributes to an environment where "technologies, people, and other entities are manipulating [interview participants'] narratives for their own ends, especially to criminalize them or their communities."

While it is unclear whether or to what extent impacted communities were given the opportunity to shape or participate in the Challenge, the NIJ was more prescriptive about who could participate through entering the Challenge as a contestant: "students; individuals/small teams/businesses; and large businesses" [4] [68]. The NIJ described those who ultimately competed in the Challenge as having a "wide variety of expertise and access to resources" [50] and explicitly centered technical expertise, stating that the Challenge "encouraged data scientists from all fields to build upon the current knowledge base for forecasting recidivism while also infusing innovative methods and new perspectives" [51]. A technosolutionist framing undercuts several of the winning entries – for example, one winner awarded $15,000 for their participation in the Challenge titled their entry "Skynet is Alive and Well: Leveraging a Neural Net To Predict Felon Recidivism" and concluded that "It is self-evident that the usage of deep learning tools has immense potential in assisting parole officers and policymakers in the rehabilitation of formerly incarcerated members of society" based on an accuracy of 67% using a neural network [45, p. 5]. Yet it is clear that some of the Challenge's winners [6, 61] lacked a basic understanding of key facets of the U.S. criminal legal system's structure and the operation of probation and parole systems. For example, one of the Challenge winners, a team that described themselves as a "group of engineers...not familiar with the recidivism literature" but with "good experience with modelling due to our work" stated that their approach was informed by their belief that "the convict has the choice to come back to jail" [6]. This team received $6,000 as a prize for their participation in the Challenge [69].

It is unclear how winning entries that display a lack of understanding of how the criminal legal system works in the United States serve the NIJ's stated goal for the Challenge of "improving outcomes for those serving a community supervision sentence" [68]. Bao et al. [8] highlight the myriad issues created by the participation of the ML community in work related to risk assessments without a clear understanding of the criminal legal context, from seemingly high quality statistical methods that actually have no impact or have harmful impacts given the realities of the context, to projects that so deeply misunderstand the criminal legal context that they view

[4] NIJ has noted that only two students entered the Challenge, one for the one year prediction window and one for the three year prediction window [70].

being rated as "high risk"[5] as the most preferable outcome of a risk assessment for a person involved with the criminal legal system [8, p. 8]. Context matters, and understanding context is not at odds with encouraging "open competition" or "making it possible for a diverse pool of entrants" to be involved [50]. Indeed, as Abebe et al. [1] highlight, technologists can sometimes play important roles in advancing social change, but having this kind of impact requires careful specification of the role of the technologist and their work, taking into account the context of the issue at hand and avoiding efforts that advance techno-solutionism [66]. As Green [41] also highlights, data scientists must recognize and engage with the fact that their work is not neutral or objective, and may be inherently political.

3.3.2 Impacts of Participation. We also examine the impacts of participation in the Challenge, asking, "What do participants own and how do they benefit?" (from [11]). Similar to Sloane et al. [84]'s recommendation that participation be recognized as work, we posit that the participation of impacted communities (e.g., through the use of one's data, including without their knowledge or permission) should be recognized as work, treated with respect, and accompanied by concrete protections [74].

While it is unclear whether the people whose data was provided to the contestants were compensated for the use of their data — or whether they were given any option to opt out of the initial data collection or subsequent use for the Challenge — the winners of the Challenge were awarded tens of thousands or even hundreds of thousands of dollars in prize money [50]. Several of these winners built models with data potentially affected by look-ahead bias, demonstrated a lack of understanding of the criminal legal system in their submissions, and exploited the Challenge's metrics, yet were still deemed winners by the NIJ and awarded large sums of money. The structure of the Challenge, including the metrics used to evaluate winners and the financial prizes based on performance along those metrics, arguably incentivized the development of solutions that could not ethically or realistically be implemented in practice for several reasons. At least two winners [18, 25], one of which was awarded $75,000 in prize money,[6] highlighted potential data leakage affecting the Challenge (through look-ahead bias), with [18] declaring that "[i]t is [their team's] conviction that any top-tier model in the Challenge will explicitly or implicitly owe its predictive prowess to the leaked information...we do not recommend the field rely on any Challenge models for future decision making" [18, p. 5]. We discuss both of these issues in more detail later in our analysis.

In addition, perhaps down- or upstream of the Challenge itself — but nonetheless important for understanding the context in which the Challenge operates — are the financial incentives and systems that encourage and benefit from the surveillance, punishment, and incarceration of poor people and people of color. Systems of parole supervision and monitoring, alongside many other parts of the criminal legal system, have clear monetary incentives and impacts, which often harm the people those systems purport to help. For in depth discussions of these incentives and their harms, see [35, 53, 54, 89, 96].

3.3.3 Contesting Participation. We analyzed if, and to what extent, challenge contestants and impacted communities had the opportunity to critique their participation in the Challenge by asking two questions: (i) "What mechanisms are put in place to allow contestants and impacted communities to question the existence of the product/tool itself, rather than purportedly helping to reduce harms or improve benefits?" and (ii) "Did contestants and impacted communities have the opportunity to refuse participation or to withdraw from the process without causing direct or indirect harm to themselves or their communities?" Both questions draw from a conceptualization of participation that comes from Birhane et al. [11].

In order to understand the contours of contesting participation, we use the definitions of the different kinds of participation (organizers, contestants, impacted communities), explained in 3.3.1. We use these different definitions of participation to start our discussion on contesting participation because we posit that an individual's level of agency in contesting the Challenge is associated with what category of participant they fall into. Contestants with winning models were able to use the conclusion, recommendations, or further consideration sections of their papers as an opportunity to comment on how the methods used in submitted models were unsuitable for use in practice, either because of the complexity of the model developed or because of strong ethical objections due to shortcomings with the provided data and evaluation metrics used for the Challenge [22, 56, 67, 92, 98, 100]. With regard to the harm a contestant would suffer by critiquing the Challenge, it appears to be primarily financial – their refusal to engage in the Challenge at all could impact their ability to reap any of the financial benefits that come with engaging in the Challenge.

For impacted communities, there is no indication in the description of the Challenge about whether they had the opportunity to contest the creation and legitimacy of the Challenge in the first place [68]. While impacted communities were not involved in the model development process, they are directly implicated in the work conducted for the Challenge. Their data was used for the Challenge and the purported impetus for the creation of the Challenge was to create models that could analyze the behavior of impacted people and enable organizers to use the models that the contestants develop to better understand and predict the behaviors of the impacted communities.

3.3.4 Depth of Participation. We center our last critique of the Challenge around two questions that interrogate the depth of participation in the Challenge: (1) Is contestant and impacted community feedback "integrated into the default ML lifecycle"? (adapted from [10]) and (2) "Will the participatory effort be a one-off engagement with the community, or a recurring/long-term engagement?" (from [11]). There is no mention made on the Challenge website about how feedback about the Challenge or issues with its construction could be made known to the Challenge organizers. The only stage, as mentioned in 3.3.3, where there would be integration of any feedback is after the winners submitted their resulting papers. There is

[5] In many criminal legal system contexts, being labelled as "high risk" by a risk assessment tool can result in punitive measures, such as additional supervision or pretrial detention.

[6] From the NIJ website, it appears that one individual was part of two winning teams — one comprised of themself (Team TrueFit [26]) which won $140,000, and the other comprised of themself and one other person (Team Crime Free) [25], which won $75,000 [69].

also no indication on the website that provides a summary of the winning papers about how the results of the Challenge would be used in practice [51].

From the design of the Challenge, it is clear that this was a one-off participation expectation — once winners submitted their models and companion research papers, their engagement with the Challenge was nearly completed. The NIJ did note that winning contestants could be asked to present their results at a speaking engagement but it is unclear whether further changes to submitted models would be required after prizes were awarded [50]. This short lifecycle of the expected period of participation may have impacted the level of care that some contestants took in developing their submitted models. Several of the winning papers offered little to no substantive critique of the Challenge, provided short summaries of their participation in the Challenge, and/or made no reference to the literature or domain that their work was situated within. One of the winners of the Challenge, [63], even blatantly acknowledged their intention to use the Challenge, not in furtherance of the creation of better risk prediction algorithms in the criminal legal system, but rather to develop a Python library, with the authors of the paper noting "the majority of the time spent on this project was devoted to the development of said library, and very little was done with the specific dataset provided by NIJ" [63]. This team was awarded $6,000 for their submission [50].

3.4 Downstream Effects of Participatory Design Failures

The failure of the Challenge organizers to use effective participatory design principles incentivized winning models that contained several methodological issues. The methodological issues that we highlight include (1) the choice of a faulty outcome variable, (2) the use of an ineffective fairness metric, (3) the inclusion of variables that could create potential feedback loops, (4) and the emergence of data leakage concerns. The methodological issues raised in the following sections are not an exhaustive list of these kinds of issues with the Challenge.

3.4.1 Choice of a Faulty Outcome Variable. In designing predictive algorithms, analysts "must translate an abstract, often ill-defined goal into a highly specified outcome variable to be predicted by the algorithm" [59]. Often, there is a large gap in the on-the-ground reality between the predictive goal and a specified outcome variable. Here, the organizers asked contestants to accurately predict recidivism which was described as "a person's relapse into criminal behavior" [71]. This definition misleadingly suggests that a person described as recidivating has actually committed a new crime. The use of arrest or re-arrest as a metric for recidivism has been widely criticized by scholars and advocates because "[o]verwhelming research has demonstrated that arrests are more reliably a measure of policing practices and priorities than actual crime" [40].

Risk assessment tools that are trained on arrest and other policing data, without evaluation of the policing culture and context in which arrests occur and data is collected, obscure the subjectivity inherent in arrest incidents [77]. The reality is that law enforcement officers are granted broad and unchecked discretion to make arrests — they decide which people to approach, which circumstances to investigate, when to document an incident as a "crime", and when to arrest people [77]. There is ample evidence that racial minorities tend to face a higher risk of arrest, especially for crimes targeted through "proactive policing" [32, 36, 94]. For example, while Black people comprise approximately 33% of the population in Georgia [16], they comprise 64% of arrests in Georgia [82], and are 1.7 times more likely than their white counterparts to be killed by the police in Georgia [82]. Where there is proactive policing over a particular group of people (in the Challenge, those subject to parole supervision), "individuals with the same probability of re-offense may nevertheless have different probabilities of re-arrest. Risk assessment tools trained on such data may appear to be fair predictors of re-arrest but nevertheless be unfair, even by the same metrics, were they to be assessed on re-offense" [36]. Additionally, information "relevant to the integrity of crime data, such as police misconduct," is not accounted for in evaluating the validity of arrest data. Questions about whether arrests were made as a result of police misconduct, or whether the circumstances of the arrests are included in the data set, are seldom posed when using arrest data [36]. Further, questions about whether departmental policies encouraged or incentivized certain arrests [15, 34, 37] are also not accounted for when considering arrest data.

Failure to consult with diverse subject matter experts and impacted communities, including people whose data is collected in the system, people and communities that are impacted by that collection, and people who can shed light on the socio-political, historical, and on-the-ground context in which data is collected, can result in selection of unrepresentative and biased outcome and input variables. Rearrest is often chosen as the outcome metric of interest to measure recidivism because it is easier to acquire arrest data than conviction data. But, "pursuing a particular outcome variable for the sake of convenience carries with it a greater risk of mismatch between the predictive goal and the variable's specification" [59]. This risk is particularly egregious when the consequences to people's lives are severe and devastating, as is the case when people are assessed as "high risk" for the purposes of release, parole, and probation supervision.

3.4.2 A Flawed Fairness Metric. In the construction of the Challenge, the NIJ sought to have some subset of the Challenge contestants develop models that took racial fairness into account. As mentioned in Section 3.1, the way that the NIJ measured fairness was with a composite fairness measure that took both the ratio of false positives between Black and white parolees and the Brier Score into account, like so: $(1 - BrierScore)(FP)$. The threshold to convert predicted probabilities to binary (yes/no) predictions was set at 0.5, so that a parolee's likelihood of recidivating had to be higher than 50% for them to be predicted to recidivate. Typically, in the machine learning model development process, the threshold to convert predicted probabilities to predicted values is chosen through experimentation with the dataset at hand [30]. However, in the case of the Challenge, the Challenge creators reported no justification for this seemingly arbitrary threshold [68]. It should be noted that there is no one "accurate" threshold for determining fairness and that the choice of a threshold can reflect subjective judgements about what a fair distribution of an outcome of interest across selected sub-populations should look like. The problem with setting this threshold arbitrarily, as several of the winning papers

noted [27, 43, 56, 76, 97, 98], was that it resulted in almost no false positives being detected. Because very few false positives were detected using the threshold set by the creators of the Challenge, contestants noted that there were no racial disparities found in the dataset, even when this is likely not the case [36, 46]. As one winning paper noted, the "fairness penalty was not very meaningful in preventing racial bias, as there were almost never any false positives for either black or white parolees" [43]. Designing a challenge that allowed for more iterative and consistent incorporation of participant feedback in the structure of the Challenge could have surfaced the critique about the threshold selected for measuring fairness, given how prevalent this critique was [27, 43, 56, 76, 97, 98], and appropriate adjustments to the Challenge could have been made so that the fairness exercise might have been fruitful. This is one of several examples where the NIJ's failure to use robust participatory design practices led to the development of models with little practical utility. While the Challenge contestants would likely not suffer the consequences of these design oversights, people upon whom a developed model would be deployed would potentially be subject to have their fates determined, in part, by a faulty algorithm.

3.4.3 Feedback Loops. Another methodological issue with the Challenge is the inclusion of features that could allow a feedback loop to be embedded into a model. Feedback loops occur when the output from a predictive model is used as input back into the model [2, 33]. Ensign et al. [33] describe the presence of feedback loops in the predictive policing context. Some jurisdictions use algorithms to decide how to allocate policing resources. Police assigned to these algorithmically-selected neighborhoods then discover activity that might not have otherwise been reported by members of the community [33]. These discovered incidents are then used as justification validating the model's predictions and are fed back into the predictive policing model as historical data, reinforcing continued police presence in existing neighborhoods, rather than reflecting true levels of crime in a particular neighborhood [33].

In the Challenge, contestants were asked to predict which parolees are likely to be re-arrested for a felony or misdemeanor. However, before the re-arrest incident happens, parolees are already tagged as having a "higher" or "lower" risk of recidivating, based on the parole supervision risk score and parole supervision level they received when they were initially released on parole. The parole supervision risk level and risk score, as documents received from the Georgia Department of Community Supervision pursuant to a FOIA request submitted by our team in 2022 found, are built on a separate risk assessment tool[7] that was designed to also predict the likelihood of arrest for a new crime during the supervision period. Crucially, the parole supervision level that parolees are assigned to influences the level of surveillance and monitoring they will be subjected to, with individuals assigned to higher levels of supervision being subjected to more frequent interactions with their parole officer. The increased frequency of interactions with parole officers, similar to the predictive policing example mentioned above, can lead to more discovered incidents and, consequently, a higher chance of re-arrest.

[7]For more information about this risk assessment tool, known as the Unified Supervision Risk Assessment Tool, please see 6.2 and 6.3

Therefore, the recidivism risk assessment instrument that the NIJ asked contestants to build already has features that could indicate likelihood of re-arrest. These features are unlikely to provide an accurate measure of riskiness but do provide an indication of the level of surveillance an individual is subjected to. Any model built off the data provided for the Challenge will likely show a strong linkage between being on a higher level of parole supervision and likelihood of being re-arrested. In essence, the NIJ asked contestants to predict an outcome that was already predicted in two of the provided features. Because a higher parole supervision level assignment results in higher surveillance, and thus more opportunities for re-arrest, individuals on higher levels of parole supervision could be tagged as more likely to recidivate which would further legitimize the use of this feature in a model, despite this rationale being flawed.

3.4.4 Data Leakage. Another potential issue with the Challenge is the possibility of data leakage. One of the winning entries to the Challenge stated that they believed they identified leakage in the data used for the Challenge, potentially compromising the Challenge and models entered to the Challenge [18]. This winner was able to identify relationships using several variables related to employment and drug testing that could be used to discern re-arrest outcomes. They suspected that look-ahead bias was potentially responsible for some of the leakage (they wrote that "Our assumption is that the values for the supervisory variables reflect their status at the time the dataset was created, i.e., sometime in 2020, instead of relative to a parolee's release date" [18, p. 20]). Furthermore, they wrote, "The preeminence of leakage is apparent, and this is the primary reason for our skepticism of any insights gained from submitted models. It is SAS' [the team that authored this submission] conviction that any top-tier model in the Challenge will explicitly or implicitly owe its predictive prowess to the leaked information. As the leakages can deterministically delineate re-arrests in a significant number of cases, no model can accurately estimate the role other variables play in affecting recidivism. Given this, we do not recommend the field rely on any Challenge models for future decision making" [18, p. 5]. At least one other winner noted the potential for some form of data leakage in the Challenge, writing that "The final two rounds may have included 'beyond the decimal point' precision in some items that provided models with some clue of when recidivism occurred (e.g. values indicating a large number of drug tests would show a parolee remained crime-free for longer)" [26, p. 8]. In this winner's analysis of feature importance, several of the most important variables in the final model were ones that Carroll et al. [18] suspected were contaminated by look-ahead bias [26, p. 7]. In its analysis "contextualizing the results" of the Challenge, the NIJ makes no mention of this potential data leakage [51], and to our knowledge, no explanation or discussion of this issue has been presented publicly by the NIJ. The fact that the Challenge may have been so severely impacted by data leakage as to potentially invalidate winning entries — by some winners' own admissions — is yet another indication of the serious methodological issues at play with the Challenge.

4 CONCLUSION

The NIJ created the Recidivism Forecasting Challenge in a stated attempt to increase public safety [68] through the development of more "innovative" [68] approaches to building recidivism risk assessment tools. Instead, the Challenge resulted in the creation of problematic predictive tools that were (1) trained on a dataset riddled with data quality issues (data leakage, feedback loops, and socio-historical biases in arrest practices), (2) developed by people who may have had little to no knowledge of the criminal legal system, (3) evaluated using a faulty threshold that surfaced effectively no disparities in the training dataset, despite known racial disparities in arrest data [36, 62, 77], (4) appeared to offer no clear documented opportunity for impacted people to critique the creation of the models or remove their personal information from the data used to build the models, and (5) created a financial incentive structure that may have encouraged the haphazard development of candidate predictive models. As highlighted in section 3 of the paper, the designers of the Challenge failed to follow guidelines for what robust participatory design could look like, resulting in models that even challenge contestants acknowledged were not suited for real world use.

Beyond influencing decisions about imprisonment and government surveillance, the data produced by law enforcement agencies and the predictions generated from risk assessment tools are often used in making decisions that can have a catastrophic impact on people's lives — including loss of parental rights, homelessness, prolonged job insecurity, immigration consequences (including deportation), and inability to access credit [5, 17, 72, 73]. As the research, development, and evaluation agency of the U.S. Department of Justice [68] tasked with investing in scientific research meant to "serve the needs of the criminal justice community," [68] the NIJ has the opportunity and responsibility to influence how local community corrections agencies think of recidivism and how to measure risk. Indeed, the Challenge website itself states that "[r]esults from the Challenge will provide critical information to community corrections departments" [68]. When the NIJ treats faulty risk assessment tools like those developed for this Challenge as "winners" deserving hundreds of thousands of dollars, the agency risks inspiring the propagation of ineffective and biased tools in local community corrections departments across the country.

Participatory design principles as well as the guidelines set out by the Blueprint and the Order provide a starting point for meaningful protections that federal, state, and local governments could implement. By involving impacted communities in decision-making and rejecting the use of biased data and faulty fairness metrics, the NIJ can design participatory processes that achieve their goal of "improving outcomes for those serving a community supervision sentence" [70]. These processes may conclude that the costs of risk assessment models far outweigh any benefit. Our analysis of the Challenge through a participatory design framework can serve as a valuable case study for future challenges of this nature. For example, several AI organizations, large tech companies, and government entities recently conducted the "largest ever public Generative AI Red Team" [64, 91], extending the established practice of assessing the security of computer systems through simulated adversarial testing to AI systems. The initiative's announcement drew on elements of participation (stating that it would "bringing in hundreds of students from overlooked institutions and communities" [91]) and noted that it is "aligned with the goals of the Biden-Harris Blueprint for an AI Bill of Rights" [91]. Our analysis could serve as a case study for this and similar efforts and could encourage future endeavors of this nature to create opportunities for meaningful and lasting participation for marginalized communities.

5 RECOMMENDATIONS FOR PRACTITIONERS

Throughout this work, we have highlighted issues with the design, execution, and results of the NIJ's Recidivism Forecasting Challenge. Grounded in the principles of participatory design, our analysis surfaced many critical questions that, had they been seriously considered by the NIJ and the Challenge contestants prior to the Challenge's development and during the Challenge itself, may have led to different outcomes. We highlight these questions in 6.1, which we intend to be a non-exhaustive list of critical questions that any government entity, individual, or organization involved in these kinds of challenges should seriously consider and answer before designing, funding, or otherwise participating in such efforts. We frame our guidance in the context of possible claims or assumptions that may underpin these kinds of efforts, using direct quotes from the Challenge's project description website as examples.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871

[2] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. 2020. Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation. In *Proceedings of the 5th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 126)*, Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.). PMLR, Virtual, 710–731. https://proceedings.mlr.press/v126/adam20a.html

[3] General Services Administration. 2023. Challenge.Gov. challenge.gov

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] American Bar Association. 2018. Collateral Consequences of Criminal Convictions: Judicial Bench Book. https://www.ojp.gov/pdffiles1/nij/grants/251583.pdf

[6] Team Aurors. 2022. National Institute of Justice Recidivism Challenge Report: Team Aurors. https://www.ojp.gov/pdffiles1/nij/grants/305053.pdf

[7] Yukino Baba, Nozomi Nori, Shigeru Saito, and Hisashi Kashima. 2014. Crowdsourced data analytics: A case study of a predictive modeling competition. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*. Institute of Electrical and Electronics Engineers, Shanghai, China, 284–289. https://doi.org/10.1109/DSAA.2014.7058086

[8] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2022. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. arXiv:2106.05498 [cs.CY]

[9] Jeffrey Bellin. 2022. *Mass Incarceration Nation How the United States Became Addicted to Prisons and Jails and How It Can Recover.* Cambridge University Press, Cambridge, United Kingdom, Chapter Distinguishing The Criminal Justice and Criminal Legal Systems, 24–30. https://doi.org/10.1017/9781009267595.006

[10] Stella Biderman and Walter J. Scheirer. 2020. Pitfalls in Machine Learning Research: Reexamining the Development Cycle. arXiv:2011.02832 [cs.LG]

[11] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. , 8 pages.

[12] Avrim Blum and Kevin Stangl. 2019. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? arXiv:1912.01094 [cs.LG]

[13] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards AI for social good. , 425–436 pages.

[14] Erica Bryant. 2021. Why We Say "Criminal Legal System," Not "Criminal Justice System. https://www.vera.org/news/why-we-say-criminal-legal-system-not-criminal-justice-system#:~:text=Lawmakers%20and%20media%20often%20speak,corrections%20in%20the%20United%20States

[15] Consumer Financial Protection Bureau. 2022. CFPB Report Shows Criminal Justice Financial Ecosystem Exploits Families at Every Stage. https://www.consumerfinance.gov/about-us/newsroom/cfpb-report-shows-criminal-justice-financial-ecosystem-exploits-families-at-every-stage/

[16] United States Census Bureau. 2023. Quick Facts – Georgia. https://www.census.gov/quickfacts/fact/table/GA/HSG651221#HSG651221

[17] Ames Grawert Cameron Kimble. 2021. Collateral Consequences and the Enduring Nature of Punishment. https://www.brennancenter.org/our-work/analysis-opinion/collateral-consequences-and-enduring-nature-punishment

[18] Mary Beth Carroll, Rodney Carson, Mike Clark, Adam Cottrell, Jim Georges, Tyler Nelson, Hiwot Tesfaye, Halil Toros, and Sree Vuthaluru. 2022. Accounting for Racial Bias in Recidivism Forecasting, Year 3 Male Parolees Report, SAS Institute Inc. Team. https://www.ojp.gov/pdffiles1/nij/grants/305056.pdf

[19] Alan Chan, Chinasa T Okolo, Zachary Terner, and Angelina Wang. 2021. The limits of global inclusion in AI development.

[20] Alex Chohlas-Wood. 2020. Understanding risk assessment instruments in criminal justice. https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/

[21] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524 [stat.AP]

[22] Giovanni Circo and Andrew Wheeler. 2022. National Institute of Justice Recidivism Forecasting Challenge: Team "MCHawks" Performance Analysis. https://www.ojp.gov/pdffiles1/nij/grants/305050.pdf

[23] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 [cs.CY]

[24] Paul Guerin Cristopher Moore, Elise Ferguson. 2023. How Much Risk, and Risk of What? A Closer Look at Pretrial Rearrest and Risk Assessment.

[25] Russell D. Wolfinger David Lander. 2022. Recidivism Forecasting with Multi-Target Ensembles: Winning Solution for Male, Female, and Overall Categories in Year One, Team CrimeFree. https://www.ojp.gov/pdffiles1/nij/grants/305032.pdf

[26] Russell D. Wolfinger David Lander. 2022. Recidivism Forecasting with Multi-Target Ensembles: Years One, Two and Three, Team TrueFit. https://www.ojp.gov/pdffiles1/nij/grants/305048.pdf

[27] Sara Debus-Sherrill and Colin Sherrill. 2022. Predicting Recidivism with Neural Network Models. https://www.ojp.gov/pdffiles1/nij/grants/305038.pdf

[28] Fernando Delgado, Solon Barocas, and Karen Levy. 2022. An uncommon task: Participatory design in legal AI. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–23.

[29] Sarah L. Desmarais, Kiersten L. Johnson, and Jay P. Singh. 2016. Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services* 13, 3 (2016), 206–222. https://doi.org/10.1037/ser0000075

[30] M. Drakesmith, K. Caeyenberghs, A. Dutt, G. Lewis, A.S. David, and D.K. Jones. 2015. Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data. *NeuroImage* 118 (2015), 313–333. https://doi.org/10.1016/j.neuroimage.2015.05.011

[31] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209. https://doi.org/10.1177/0093854818811379 arXiv:https://doi.org/10.1177/0093854818811379

[32] Cindy Reed Elizabeth Hinton, LeShae Henderson. 2018. An Unjust Burden: The Disparate Treatment of Black Americans in the Criminal Justice System. , 19 pages. https://www.vera.org/downloads/publications/for-the-record-unjust-burden-racial-disparities.pdf

[33] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. *CoRR* abs/1706.09847 (2017), 1–12. arXiv:1706.09847 http://arxiv.org/abs/1706.09847

[34] Jackie Fielding. 2022. Outlawing Police Quotas. https://www.brennancenter.org/our-work/analysis-opinion/outlawing-police-quotas

[35] Fines and Fees Justice Center. 2022. Electronic Monitoring Fees: A 50-State Survey of the Costs Assessed to People on E-Supervision. https://finesandfeesjusticecenter.org/content/uploads/2022/09/FFJC-Electronic-Monitoring-Fees-Survey-2022.pdf

[36] Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. arXiv:2105.04953 [stat.AP]

[37] Brennan Center for Justice. 2023. How Perverse Financial Incentives Warp the Criminal Justice System. https://www.brennancenter.org/series/how-perverse-financial-incentives-warp-criminal-justice-system

[38] National Artificial Intelligence Research Resource Task Force. 2023. Strengthening andDemocratizing theU.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

[39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. arXiv:1803.09010 [cs.DB]

[40] Marissa Gerchick and Brandon Buskey. 2022. Formal Statement of the American Civil Liberties Union For a Stakeholder Engagement Session on First Step Act Implementation. , 9 pages. https://www.aclu.org/other/aclu-statement-pattern-risk-assessment-tool

[41] Ben Green. 2021. Data science as political action: grounding data science in a politics of justice. *Journal of Social Computing* 2, 3 (2021), 249–265.

[42] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. https://doi.org/10.1007/s13347-022-00584-6

[43] Timothy Han. 2022. Recidivism Forecasting Using XGBoost. https://www.ojp.gov/pdffiles1/nij/grants/305033.pdf

[44] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[45] Dylan Hanson. 2022. Skynet is Alive and Well: Leveraging a Neural Net to Predict Felon Recidivism. https://www.ojp.gov/pdffiles1/nij/grants/305052.pdf

[46] Bernard Harcourt. 2015. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter* 27 (04 2015), 237–243. https://doi.org/10.1525/fsr.2015.27.4.237

[47] Melissa Heikkilä. 2022. A bias bounty for AI will help to catch unfair algorithms faster. https://www.technologyreview.com/2022/10/20/1061977/ai-bias-bounty-help-catch-unfair-algorithms-faster/

[48] White House. 2022. Blueprint for An AI Bill of Rights: Making Automated Systems Work for the American People. , 73 pages. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[49] White House. 2023. Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

[50] Caleb Hudgins, Veronica White, D. Michael Applegarth, and Joel Hunt. 2021. Results from the National Institute of Justice Recidivism Forecasting Challenge. https://nij.ojp.gov/topics/articles/results-national-institute-justice-recidivism-forecasting-challenge

[51] Caleb Hudgins, Veronica White, D. Michael Applegarth, and Joel Hunt. 2022. The NIJ Recidivism Forecasting Challenge: Contextualizing the Results. https://nij.ojp.gov/library/publications/nij-recidivism-forecasting-challenge-contextualizing-results

[52] Kosuke Imai, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. 2020. Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. https://doi.org/10.48550/ARXIV.2012.02845

[53] Prison Policy Initiative. 2023. Economics of Incarceration: The economic drivers and consequences of mass incarceration. https://www.prisonpolicy.org/research/economics_of_incarceration/

[54] Prison Policy Institute. 2023. Economics of Incarceration: The economic drivers and consequences of mass incarceration. https://www.prisonpolicy.org/research/economics_of_incarceration/

[55] Carrie Johnson. 2022. Flaws plague a tool meant to help low-risk federal prisoners win early release. https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

[56] Team Klus. 2022. NIJ Recidivism Challenge Report, Team Klus. https://www.ojp.gov/pdffiles1/nij/grants/305040.pdf

[57] Tammy Rinehart Kochel, David B. Wilson, and Stephen D. Mastrofski. 2011. Effect of Suspect Race on Officers' Arrest Decisions. *Criminology* 49, 2 (2011), 473–512. https://doi.org/10.1111/j.1745-9125.2011.00230.x arXiv:https://misclibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.2011.00230.x

[58] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning.

[59] David Lehr and Paul Ohm. 2017. Playing with the Data: What Legal Scholars Should Learn About Machine Learning. , 655-717 pages. https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf

[60] Robert Long. 2021. Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness. *Journal of Moral Philosophy* 19, 1 (2021), 49 – 78. https://doi.org/10.1163/17455243-20213439

[61] Yujunrong Ma, Kiminori Nakamura, Eung-Joo Lee, and Shuvra S. Bhattacharyya. 2022. National Institute of Justice's Recidivism Forecasting Challenge: Research Paper, Group MNLB. https://www.ojp.gov/pdffiles1/nij/grants/305046.pdf

[62] Sandra G. Mayson. 2019. Bias In, Bias Out. *The Yale Law Journal* 128, 8 (2019), 2218–2300. http://www.jstor.org/stable/45098041

[63] Murray Miron. 2022. National Institute of Justice's Recidivism Forecasting Challenge, SRLLC. https://www.ojp.gov/pdffiles1/nij/grants/305041.pdf

[64] Alan Mislove. 2023. Red-Teaming Large Language Models to Identify Novel AI Risks. https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/

[65] John Monahan and Jennifer L. Skeem. 2016. Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology* 12, 1 (2016), 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945 arXiv:https://doi.org/10.1146/annurev-clinpsy-021815-092945 PMID: 26666966.

[66] Evgeny Morozov. 2013. To save everything, click here: The folly of technological solutionism.

[67] Matthew Motoki and Sorapong Khongnawang Marifel Barbasa. 2022. Team MattMarifelSora: NIJ Recidivism Forecasting Challenge Report. https://www.ojp.gov/pdffiles1/nij/grants/305044.pdf

[68] National Institute of Justice. 2021. Description of the Recidivism Forecasting Challenge. https://nij.ojp.gov/funding/recidivism-forecasting-challenge

[69] National Institute of Justice. 2021. Recidivism Forecasting Challenge Awards. https://nij.ojp.gov/funding/opportunities/nij-recidivism-forecasting-challenge?page=0#awards-block-2-cvykkgn-iuzm9fuk

[70] National Institute of Justice. 2022. Recidivism Forecasting Challenge: Official Results. https://nij.ojp.gov/funding/recidivism-forecasting-challenge-results

[71] National Institute of Justice. 2023. Recidivism. https://nij.ojp.gov/topics/corrections/recidivism#:~:text=Recidivism%20is%20one%20of%20the,intervention%20for%20a%20previous%20crime.

[72] U.S. Commission on Civil Rights. 2019. Collateral Consequences: The Crossroads of Punishment, Redemption, and the Effects on Communities. https://www.usccr.gov/files/pubs/2019/06-13-Collateral-Consequences.pdf

[73] Devah Pager. 2003. The Mark of a Criminal Record. https://scholar.harvard.edu/files/pager/files/pager_ajs.pdf

[74] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[75] Tawana Petty, Mariella Saba, Tamika Lewis, Seeta Peña Gangadharan, and Virginia Eubanks. 2018. Our data bodies: Reclaiming our data. *June* 15 (2018), 37.

[76] Michael Porter and George Mohler. 2022. NIJ Recidivism Forecasting Challenge Report for Team PASDA. https://www.ojp.gov/pdffiles1/nij/grants/305042.pdf

[77] Kate Crawford Rashida Richardson, Jason M. Schultz. 2022. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice.

[78] David G Robinson. 2022. Voices in the Code: A Story about People, Their Values, and the Algorithm They Made.

[79] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[80] Meera Santhanam. 2021. Criminal Justice or Criminal Injustice? The Power of Language. https://pulitzercenter.org/stories/criminal-justice-or-criminal-injustice-power-language

[81] Ric Simmons. 2020. Big Data and Procedural Justice: Legitimizing Algorithms in the Criminal Justice System. https://doi.org/10.2139/ssrn.3659347

[82] Samuel Sinyangwe. 2023. Police Scorecard - Georgia. https://policescorecard.org/ga

[83] Mona Sloane. 2020. Participation-washing could be the next dangerous fad in machine learning).

[84] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning.

[85] John Stamper and Zachary A Pardos. 2016. The 2010 KDD Cup Competition Dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics* 3, 2 (Sep. 2016), 312–316. https://doi.org/10.18608/jla.2016.32.16

[86] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. , 1162–1177 pages.

[87] Megan Stevenson and Jennifer L. Doleac. 2020. Algorithmic Risk Assessment in the Hands of Humans. https://doi.org/10.2139/ssrn.3513695

[88] Megan Stevenson and Sandra Gabriel Mayson. 2022. Pretrial Detention and the Value of Liberty. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787018

[89] Ram Subramanian, Jackie Fielding, Lauren-Brooke Eisen, Hernandez Stroud, and Taylor King. 2022. Revenue Over Public Safety.

[90] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Ángeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. , 667–678 pages.

[91] Austin Carson Sven Cattell, Rumman Chowdhury. 2023. AI Village at DEF CON announces largest-ever public Generative AI Red Team. https://aivillage.org/generative%20red%20team/generative-red-team/

[92] Duddon Evidence to Policy Research Team. 2022. Predicting Recidivism in Georgia Using Lasso Regression Models with Several New Constructs. https://www.ojp.gov/pdffiles1/nij/grants/305049.pdf

[93] Pam Ugwudike. 2022. Predictive Algorithms in Justice Systems and the Limits of Tech-Reformism. *International Journal for Crime, Justice and Social Democracy* 11, 1 (Mar. 2022), 85–99. https://doi.org/10.5204/ijcjsd.2189

[94] American Civil Liberties Union. 2020. Marijuana Arrests by the Numbers. https://www.aclu.org/gallery/marijuana-arrests-numbers

[95] American Civil Liberties Union. 2022. ACLU COMMENT ON NIST'S SECOND DRAFT AI RISK MANAGEMENT FRAMEWORK. https://www.aclu.org/letter/aclu-comment-nists-second-draft-ai-risk-management-framework

[96] Michael Waldman. 2022. Perverse Financial Incentives in Criminal Justice. https://www.brennancenter.org/our-work/analysis-opinion/perverse-financial-incentives-criminal-justice

[97] Jeremy Walthers. 2022. Recidivism Forecasting Challenge: Team IdleSpeculation Report. https://www.ojp.gov/pdffiles1/nij/grants/305034.pdf

[98] David B. Wilson, Evan M. Lowder, Peter Phalen, and Ashley Rodriguez. 2022. National Institute of Justice's Forecasting Recidivism Challenge: Team "DEAP" (Final Report). https://www.ojp.gov/pdffiles1/nij/grants/305037.pdf

[99] Aleš Završnik. 2021. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology* 18, 5 (2021), 623–642. https://doi.org/10.1177/1477370819876762 arXiv:https://doi.org/10.1177/1477370819876762

[100] Cengiz Zopluoglu. 2022. Forecasting Recidivism: Mission Impossible. https://www.ojp.gov/pdffiles1/nij/grants/305054.pdf

[101] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric PS Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. , 4 pages.

6 APPENDICES

6.1 Recommendations for Questions to Ask When Considering Designing or Participating in Crowd-Sourced Challenges

The following is a non-exhaustive list of critical questions that any government entity, individual, or organization involved in these kinds of challenges should seriously consider and answer before designing, funding, or otherwise participating in such efforts, framed in the context of possible claims or assumptions that may underpin these kinds of efforts. The list is structured in the following way:

- **Assumptions or claims that might underpin a Challenge's design**
  (1) **Related questions that practitioners should consider**

List of Claims and Related Questions

- **"The Challenge aims to improve the ability to forecast recidivism...with the goal of improving outcomes for those serving a community supervision sentence" [68].**
  (1) How am I defining "improving outcomes," and who did I consult in developing that definition? Did I consult people under community supervision, and are their inputs reflected in this definition? Did I consult people who advocate on behalf of those under community supervision?
  (2) If my stated goal is to "improve the ability to forecast recidivism," how am I defining "recidivism," and is that an outcome that can be measured, or am I relying on proxy variables like re-arrest or return to custody?
  (3) Do the proxy variables actually predict the outcome I am interested in? Or do they offer information that is too problematic or removed such that they should not be used? Do I need outside help in answering these questions?
  (4) Have there been rigorous independent analyses assessing whether forecasting recidivism improves outcomes for people on community supervision? If so, what does this evidence indicate? If not, is this a viable goal?
  (5) What alternative interventions can improve outcomes for people on community supervision? Have I considered investing in those alternatives rather than conducting this challenge (including as part of a cost-benefit analysis)?
- **"The Challenge provided the public with open data access, making it possible for a diverse pool of entrants to compete in the Challenge" [50]**.
  (1) Do the individuals participating in the challenge understand the social, political, and historical context in which they are developing predictive tools?
  (2) What forms of knowledge are prioritized in the development of the challenge? Are you primarily engaging technical experts?
  (3) Are impacted communities and those who advocate for impacted communities afforded meaningful opportunities to participate?
  (4) How are impacted communities engaged in the ideation and development phases of the challenge? Are impacted communities given veto power throughout the lifecycle of the challenge?
  (5) Who is designated as a stakeholder in the development of the desired predictive tool? How am I defining what a "diverse" pool of entrants looks like? Do I need to seek outside help in examining my assumptions?
- **The Challenge will use a "measure of fairness" "to calculate which algorithms are the most accurate while accounting for bias" [68].**
  (1) What does the chosen fairness metric prioritize?
  (2) What assumptions about fairness are baked into the choice of evaluation metric?
  (3) Is there a way for impacted communities or advocates for impacted people to offer input about the assumptions?
  (4) Should we be building a tool that is based on biased data?
  (5) Can a fairness metric meaningfully account for bias?
- **"Findings could directly impact the types of factors considered when evaluating risk of recidivism and highlight the need to support people in specific areas related to reincarceration." [68]**
  (1) Have you considered how the challenge could provide support for non-carceral solutions to recidivism prevention? What sort of harms were considered and valued in the development of the predictive tools to determine that the findings were robust?
  (2) Have you developed pipelines for community feedback and engagement with the results of the challenge?
  (3) Have I considered how to make my findings transparent and open to impacted communities and not just those involved in the carceral apparatus?

6.2 Description of the Development of the Unified Supervision Risk Assessment Tool in Georgia

ARS
APPLIED RESEARCH SERVICES, INC.

Georgia Department of Community Supervision
Unified Supervision Risk Assessment Project Update
Tammy Meredith, Ph.D., Applied Research Services, Inc.
6/22/2017

In September 2014, we proposed to the Georgia Board of Pardons and Paroles a new study of parolees to determine the risk factors (both static and dynamic) that drive supervision success in order to update our 2010 automated supervision risk-assessment instruments. Since then, parole and probation supervision were combined under the direction of the Department of Community Supervision. Consequently, this project was restructured to include both parolees and probationers to help DCS move to a seamless "unified" system of automating the risk-assessment process with inform supervision level assignments.

The new project was approved in March 2016, to be completed within 24 months beginning April 1, 2016, with Dr. Tammy Meredith of ARS serving as Project Director. The 24-month timeline was articulated as "dependent upon the accuracy and timely receipt" of the required CMS parole and GDC Scribe data tables. While significant delays in data transfer have resulted in project interruptions, the following update describes the progress to date for Phase 1 (the first 12-months of work) and makes recommendations for refinements to the Phase 2 work plan (second 12-months of work). At this point we recommend a no-cost project extension for six months, as described in detail below.

Phase 1: Develop and Validate New Supervision Assessment Tools (12 Months 4/1/2016 – 3/31/2017)

The following deliverables were established for Phase 1, with notes provided on current status:

Deliverable	Status
1. Provide Written & Oral Updates to DCS	on-going
2. Add Probation Supervision (original study designed for parole)	completed
3. Define Study Cohorts (probation & parole) and Data Requirements	completed
4. Finalize Methodology/Analytical Plan	completed
5. Determine the Number of Instruments Needed	recommendations made to DCS
6. Address all Management Research Questions	on-going
7. Provide All Computer Algorithms to DCS IT (universal text file format)	pending analysis completion
8. Consult/Train DCS IT Representatives on New Algorithms	pending analysis completion
9. Translate Risk Scores into Scales, Set Supervision Level Cut Points, Work on Display of Results	pending analysis completion
10. Test and Implement New Tools at DCS	pending analysis completion
11. Make Recommendations for Overlap with GDC's NGA Tools	recommend move to Phase 2

Phase 1: Update

From the April 1, 2016 project start, we anticipated that data requests and secure transfer of all data tables from three agencies (DCS/Parole, GDC, and GCIC) could easily be completed within the first six months of the project (by September 2016). This timetable should have been sufficient, given ARS extensive prior experience with all three agency data systems. While GDC and GCIC data were all received in less than six months, the last DCS/parole data table arrived 13 months into the project, May 3, 2017. Most of the DCS data tables were sent multiple times, in differing formats, which added to the delays. Thus, our 12-month timeline for Phase 1 has been delayed. We anticipate completing Phase 1 deliverables #7 through #9 (see above) by July 30, 2017. That accelerated analysis timeline would leave the project Phase 1 four months behind schedule. Then, deliverable #10 will require DCS management commitment to be completed in a timely manner. Deliverable #11 will be discussed below in Phase 2.

A significant amount time has been devoted to developing a probation study cohort, comparable to the parole cohort created when the project started (with Parole). Supervision activity data tables for probationers were the most problematic in transfer and delivery, not content. Data on the study cohorts were presented to the DCS project management team in November 2016. DCS management selected the outcome of interest for the new assessment tool as "an arrest for a new crime during supervision," consistent with both existing parole and probation supervision assessment tools. The outcome is measured with GCIC official arrest records and identify arrests where the most serious offense must be a new crime, not a probation or parole violation arrest, within 24 months of starting probation.

The study cohorts (described below) include all admissions to either probation or parole supervision during calendar years 2011-2012, thus leaving sufficient time elapse to collect up to 24 months of supervision activity and potential arrest. The probation cohort is younger, has more females (21%), and on supervision longer (an average of five months); nearly half are still on supervision.

	Parole Cohort	Probation Cohort
Sample Size (# of offenders)	25,679	94,619
Started Supervision		
2011	52%	49%
2012	48%	51%
Ended Supervision		
2011	8%	4%
2012	30%	11%
2013	29%	12%
2014	14%	13%
2015-2017	10%	14%
Still Active	9%	46%
Average Years on Supervision (to date)	1.3 yrs	1.7 yrs
Demographics		
Male	89%	79%
Nonwhite	58%	57%
Average Age at Supervision Start	36 yrs	33 yrs

Deliverable #5 (Determine the Number of Instruments Needed) is the current focus of analysis. We conducted an extensive analysis of the "types" of community supervision populations portrayed in the data. Detailed profiles of probation and parole populations are required to see how they are different or similar in terms of demographics, conviction offense, prior criminal history, prison experience, and supervision behavior. If they are similar, then DCS may decide to move toward a single instrument. Most of our research in Georgia indicates that traditional probation populations are younger and have less extensive criminal histories, thus their drivers of risk (both risk factors and weights) would be different. In addition, we suspect that the "traditional" probation population in Georgia actually encompasses two subpopulations – those coming directly from court to probation, and those exiting prison to probation on a split sentence. In the second scenario, our split sentence probationers may be more similar to parolees.

Additional comparisons of the cohorts are provided below. The probation cohort is divided into those entering supervision straight from court (52%) and by existing prison on a split sentence (48%).

			Comparing 2 Probation Cohorts	
	Parole Cohort	**Probation Cohort**	**Straight from Court**	**Split/Exit Prison**
Sample Size (# of offenders)	25,679	94,619	45,822	42,856
Primary Offense Type				
Personal	28%	21%	16%	29%
Property	33%	36%	41%	34%
Drug	27%	26%	32%	23%
Prior Arrest Record (GCIC)				
Felony Offense	98%	98%	99%	99%
Misdemeanor Offense	79%	72%	67%	78%
Violent Offense	55%	44%	35%	54%
Sex Offense	6%	6%	4%	8%
Property Offense	71%	68%	66%	72%
Drug Offense	68%	57%	54%	62%
Parole/Prob Violation Offense	72%	47%	34%	62%
First Supervision Level				
Standard	43%	63%	73%	63%
High/Specialized	46%	16%	13%	16%
Contact	1%			
Admin		11%	10%	11%
Warrant		3%	1%	3%
Other		7%	3%	7%
Last Supervision Level				
Standard	49%	43%	54%	73%
High/Specialized	35%	8%	8%	13%
Contact	6%			
Admin		25%	17%	10%
Warrant		14%	12%	1%
Other		10%	9%	3%
Average # Supervision Level Changes	1.9	2.2	2.1	2.6

As anticipated, parolees and split sentence probationers share many similarities. Both are much more likely to be serving time in the community for a personal offense, compared to straight probationers. Yet differences are noted; high or specialized supervision level assignments are much more common among parolees, likely due to differences in agency policies prior to the DCS formation. However, there is more movement across supervision levels (instability) among probationers, particularly split sentence offenders.

Finally, comparisons across cohorts on the outcomes of interest are provided below, specifically highlighting arrest for a new crime during the first 24 months of supervision. Split sentence probationers are most likely to be arrested within 24 months of starting supervision (38% fail), compared to similar levels of failure among straight probationers and parolees (27%-28%).

			Comparing 2 Probation Cohorts	
	Parole Cohort	**Probation Cohort**	**Straight from Court**	**Split/Exit Prison**
24 Month Supervision Key Outcomes				
Failed a Drug Test*	28%	17%	18%	18%
Arrested for New Crime	**28%**	**31%**	**27%**	**38%**
Arrested for New Felony Crime	14%	16%	10%	23%
Revoked	17%	11%	5%	23%
* Received a Drug Test	77%	41%	42%	45%

The data suggests that a simple binary grouping of offenders, combining both groups that exit from prison onto supervision, might not be optimal. Three unique assessment instruments may be required, given the complicated profiles and differences in outcomes across our three subpopulations. To further examine this question, preliminary multivariate logistic regression statistical techniques were employed to identify the anticipated core predictors of the outcome (such as gender, race, offense type, and prior arrests).

Grouping the three subpopulations together results in statistically significant differences. For example, our preliminary models indicate that the likelihood of a new arrest is over 50% higher for split sentence compared to straight probationers. Next, our models suggest that parolees do not share the same predictive patterns as either group of probationers. Therefore, we recommend that DCS move forward with developing three unique assessment tools, one for each specific subpopulation. In addition, we recommend that gender-specific models will be most appropriate as males are significantly more likely to be arrested within each subpopulation,. Accordingly, we anticipate developing six total risk-assessment algorithms by the end of July 2017.

As outlined in our project proposal, we will divide each of the three subpopulations into two random samples of equal size to complete a validation study of the new tools. The "test" samples will be used to conduct the statistical analyses that define two predictive equations (for males and females). Then we will score the "validation" samples on the selected algorithms to examine how accurately it predicts the outcome (new crime arrest during supervision). Since we allot for 24 months of supervision completion for each study cohort, the outcomes will be known (did they get arrested within 24 months of supervision?).

Phase 2: Monitoring, Evaluation & Updates (12 Months 4/1/2017 – 3/31/2018)

The following deliverables were established for Phase 2:

Deliverable

1. Provide Quarterly Written & Oral Research Updates to the DCS Oversight Team
2. Provide Management & Personnel Training on New Tool Implementation
3. Monitor New Tool Implementation
4. Request & Examine New Data Extracts (risk factor data & DCS calculated risk scores and groups)
5. Work Closely with the DCS to Make Adjustments that Ensure Accurate Implementation
6. Examine New Data at 6 & 12 Month Intervals to Conduct Validation Studies
7. Provide DCS Analytical Assistance, Advice & Oversight on Agency Protocol Development
8. Provide a Final 24-month Written Project Report (with DCS review & input)

Phase 2: Update

Training on new tool operation and monitoring implementation will be conducted as planned, after Phase 1 tasks are completed. Once DCS (Parole) IT executes the new assessment algorithms on agency computers, we will request new data extracts from DCS and GCIC to examine and monitor performance – focusing on whether the estimates of risk on new supervision cases look similar to the development and original validation cohorts.

Given the four-month delay in the Phase 1 analysis plan, and leaving sufficient time for DCS to implement the new algorithms, we recommend a full 6-month no-cost extension of the project. The proposed revised Phase 2 timeline would be the 12-month period of 10/1/2017 – 9/30/2018. While we are confident in our ability to complete the project by the end of September 2018 (6 months behind original schedule), this timeline will require the accurate and timely receipt of DCS data tables and DCS (Parole) IT implementation. Delays requested by DCS can certainly result in adjustments to the new proposed timeline.

Given the project delays, sufficient time will not elapse after implementation of the algorithms on DCS computers to allow for more than a 6-month interval to validate the tools on a new population to assess the accuracy of arrest predictions (Deliverable #6). We will outline for DCS the requirements for true validation upon the project completion, to include a 24-month follow-up period on a new cohort of offenders.

We recommend that Phase 1 Deliverable #11 (Make Recommendations for Overlap with GDC's NGA Tools) be moved to Phase 2 for multiple reasons. First, many discussions with DCS management have resulted in new ideas for the comparison of our new DCS supervision risk assessments with the GDC NGA inmate risk/needs assessments. Second, we have discovered in our Phase 1 analyses that NGA assessments cannot be retrospectively produced for the historical study cohort. Finally, our original plan was to examine the existing DCS risk measure, defined by the existing parole and probation supervision risk assessment tools, against the NGA needs scales (since there are no comparable needs scales for DCS populations). That line of inquiry would help DCS develop risk/need matrices to inform treatment plans

and program placement. Since supervision risk has been defined for years for the DCS populations, it would be a more valid measure of risk than applying GDC-developed risk tools (NGA risk scales), which were developed primarily on inmates and exclude the majority (90%) of probationers. We have since discovered that DCS has not taken over the calculation of the supervision risk algorithm for probationers from GDC. Our study cohort includes parole supervision risk data, but no probation risk data.

That leaves two options for completing a DCS risk/needs analysis: go back to GDC and request their calculation of the probation supervision risk-assessment algorithm results (which they continue to calculate), or wait until we complete the new DCS unified risk-assessment algorithms. We would also suggest both options are possible during Phase 2.

Finally, to aid in the development of DCS risk/need matrices to inform treatment plans we have begun analysis on the current DCS population. A description of that cohort is provided below.

	Total	Parolees	Probationers	Parole & Probation Dual Supervision
Active Population March 2017	229,617	19,172	206,246	4,168
Started Supervision				
up to 2007	8%	10%	8%	14%
2008	3%	2%	3%	4%
2009	4%	3%	4%	5%
2010	5%	4%	5%	6%
2011	7%	5%	7%	7%
2012	10%	5%	11%	11%
2013	12%	8%	12%	12%
2014	15%	12%	15%	16%
2015	17%	18%	17%	15%
2016	18%	30%	17%	10%
2017	1%	3%	1%	0%
Average Years on Supervision (to date)	4.4 yrs	4.1 yrs	4.4 yrs	5.7 yrs
Demographics				
Male	79%	90%	78%	88%
Nonwhite	54%	63%	54%	54%
Average Age at Start	34 yrs	38 yrs	34 yrs	34 yrs
Primary Offense Type				
Personal	19%	13%	21%	16%
Property	27%	15%	30%	30%
Drug	23%	18%	25%	18%
Current Supervision Level				
Unsupervised	12%	0%	13%	1%
Contact	6%	8%	6%	3%
Standard	39%	61%	37%	38%
High	5%	11%	5%	23%
Special	5%	7%	4%	4%
DRC	1%	1%	1%	1%
Warrant	17%	8%	18%	13%
Other/Old Codes	15%	4%	16%	17%

The first step in building a risk/needs matrix is a valid (accurate) measure of risk, which for DCS is defined as a new crime arrest during supervision. GDC has provided all existing NGA risk and need scale data for the current DCS cases, which covers 80% of the active population. That proportion differs by subpopulation with long term parolees having the least amount of data (since their cases began prior to the 2014 start of NGA) as described below.

	Total	Parolees	Probationers	Parole & Probation Dual Supervision
Active Population March 2017	229,617	19,172	206,246	4,168
% Cases with DCS Supervision Risk Data	15%	98%	6%	99%
% Cases with NGA Risk Data	81%	59%	83%	87%
% Cases with NGA Need Data	81%	59%	83%	87%

However, the NGA scales were neither developed nor validated on standard probationers. Since DCS does not have risk data for probationers, our recommendation is to wait for the new algorithms from Phase 1 to use as the new measure of risk.

The second step in building a risk/needs matrix is a valid measure of offender needs, or those malleable risk factors that can be reduced through evidence-based programming (the targets of intervention). We recommend focusing Phase 2 analysis time on testing the GDC NGA need scales on the DCS population to determine whether data exists to populate the algorithms, and whether the algorithms are predictive among the DCS population. Such an analysis will allow us to determine which factors on the NGA needs scales are valid for the DCS population, what additional data may be required to supplement NGA measures of need, and how to develop a DCS-appropriate risk/needs matrix. We recommend this line of analysis in Phase 2, which we propose to complete by September 30, 2018.

Unified Supervision Risk Assessment Project Team

Applied Research Services, Inc.
Tammy Meredith, Ph.D.
Shila Réne Hawk, Ph.D.
(404) 881-1120
meredith@ars-corp.com
shawk@ars-corp.com

Georgia Department of Community Supervision
Sherri Bloodworth, Field Services
(404) 309-6346
sherri.bloodworth@dcs.ga.gov

6.3 Georgia Department of Community Supervision Unified Risk Assessment Features & Weights

Georgia Department of Community Supervision Unified Risk Assessment 2017

T. Meredith, Ph.D., Applied Research Services 9/12/2017

2017 Unified Supervision Risk Algorithm for Parole Releases from Prison

Males

Enter the answers for your case in "Score" column to compute offender's risk probability

Risk Factor	Weight	Score	WT*Score	Score Instructions
Age at Supervision Start	-0.042	41	-1.722	enter age in years
Primary Offense is Property (Y/N)	0.216	1	0.216	enter 0 (no) or 1 (yes)
Prison Admission was for Revocation (Y/N)	0.402	0	0	enter 0 (no) or 1 (yes)
# of Prior Prison Episodes	0.113	6	0.678	enter # of prior prison episodes
# Prior Arrest Episodes	0.049	25	1.225	enter # of prior arrest events
# of Prison Disciplinary Reports	0.046	0	0	enter # of prison DRs
Mental Health Problem (Y/N)	0.167	1	0.167	enter 0 (no) or 1 (yes)
Prison STG Validation (Y/N)	0.314	0	0	enter 0 (no) or 1 (yes)
# Supervision Violations - Moving	0.139	6	0.834	enter # of violations for moving
# Supervision Violations - Alcohol	0.304	0	0	enter # of violations for alochol
# Supervision Violations - Violence	0.452	0	0	enter # of violations for violence
# of Positive Drug Screens	0.121	0	0	enter # of positive drug screens
# Program Removals	0.167	2	0.334	enter # of program removals
Constant	-0.677			
Logit			1.055	
Probability of Felony Arrest			0.742	

PRELIMINARY/TEST Risk Group	Low End Probability	High End Probability
Low	0.000	0.250
Medium	0.250	0.382
High	0.382	1.000

Enter the answers for your case in "Score"

Females			*column to compute offender's risk probability*	
Risk Factor	Weight	Score	WT*Score	Score Instructions
Age at Supervision Start	-0.033	27	-0.891	enter age in years
Primary Offense is Property (Y/N)	0.341	0	0	enter 0 (no) or 1 (yes)
# Prior Arrest Episodes	0.062	13	0.806	enter # of prior arrest events
# of Prison Disciplinary Reports	0.045	2	0.09	enter # of prison DRs
Mental Health Problem (Y/N)	0.314	1	0.314	enter 0 (no) or 1 (yes)
# Supervision Violations - Moving	0.159	1	0.159	enter # of violations for moving
# Supervision Violations - Violence	1.368	0	0	enter # of violations for violence
# of Positive Drug Screens	0.166	0	0	enter # of positive drug screens
# Program Removals	0.241	5	1.205	enter # of program removals
Constant	-1.743			
Logit			-0.060	
Probability of Felony Arrest			0.485	
	Low End	High End		
PRELIMINARY/TEST Risk Group	Probability	Probability		
Low	0.000	0.165		
Medium	0.165	0.276		
High	0.276	1.000		