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

NATIONAL INSTITUTE OF JUSTICE

JANUARY 2025

A SYNTHESIS OF THE 2021 NIJ FORECASTING CHALLENGE WINNING REPORTS

By Veronica White, Rachel Rief, Caleb D. Hudgins, Meaghan L. Pimsler, Joel Hunt

NIJ | NATIONAL INSTITUTE OF JUSTICE
ADVANCING JUSTICE THROUGH SCIENCE

U.S. Department of Justice
Office of Justice Programs
999 N. Capital St. N.E.
Washington, DC 20531

Nancy La Vigne, Ph.D.
Director, National Institute of Justice

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

National Institute of Justice
Advancing Justice Through Science
NIJ.ojp.gov

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

The National Institute of Justice is the research, development, and evaluation agency of the U.S. Department of Justice. NIJ's mission is to foster and disseminate knowledge and tools derived from objective and rigorous scientific research to inform efforts to promote safety and advance justice.

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

NCJ 309826

Table of Contents

Introduction

The National Institute of Justice (NIJ) hosted the 2021 Recidivism Forecasting Challenge (hereafter referred to as the Challenge), which evaluated team forecasts of the probability of individuals on parole recidivating during a specified time interval.[1] NIJ anticipated that the Challenge would inspire teams to apply innovative data science techniques to forecast the probability that individuals would recidivate and inherently further our general knowledge of what variables are important in forecasting recidivism. (See https://nij.ojp.gov/funding/recidivism-forecasting-challenge for a full description of the structure of the Challenge, including the background, rationale, description of who is in the dataset, variables, team/contestant types, and prize categories.)

Submissions were evaluated based on two different statistics: the Brier score and a Fairness score. A Brier score is a method of evaluating the accuracy of forecasts. In technical, mathematical language, a Brier score is the average squared error in forecasts.[2] In less technical terms, it is the average difference between the forecasted probability and the actual outcome. A Brier score, therefore, falls between 0 and 1, where the lower the error (and therefore the Brier score), the more accurate the forecast. Three individual Brier scores were computed: males in the dataset, females in the dataset, and the average of these two scores.

The second statistic looked at the racial fairness of the forecasts and was called the Fairness score. This metric incorporated the difference in false positive rates (forecasted to recidivate but did not recidivate) between Black and white individuals on parole. The difference in false positive rates was used as a penalty that was applied to the Brier score. This was computed for males in the dataset and females in the dataset; no average was calculated. (See Judging Criteria in the Challenge description for the equations used. Additional background on these metrics and a short analysis of how well winning submissions performed based on these two criteria can be found in Results from the National Institute of Justice Recidivism Forecasting Challenge[3]).

To add to the overall body of knowledge of risk assessment creation, NIJ required the winning teams to submit a short paper describing (1) what type of model/algorithm(s) was used, (2) what variables mattered in their model (when possible), (3) additional variables created or added to their model that NIJ did not provide, (4) if/how models/variables differed between genders, and (5) if/how they accounted for any potential racial bias in their forecasting model, and other key findings. This paper aims to add to the knowledge of risk assessment creation by synthesizing the 25 winning, non-student papers (see winning paper submissions).[4]

[1] For this paper, forecasts refer to the 0/1 predictions of recidivism. Prediction is used to refer to the (percent) likelihood of recidivism.

[2] Glenn W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78 no. 1 (1950): 1-3, https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

[3] "Recidivism Forecasting Challenge," *National Institute of Justice*, https://nij.ojp.gov/funding/recidivism-forecasting-challenge.

[4] NACJD page for all winning paper submissions: https://nij.ojp.gov/topics/articles/results-national-institute-justice-recidivism-forecasting-challenge#papers-from-the-winners.

The Importance of Fairness and Bias in Risk Assessments

At the onset of this paper, we want to highlight the importance of fairness and bias in risk assessments. During the drafting of the competition, which included the requirements of winners' reports, we did not stipulate that the reports needed to include a discussion on the importance of fairness or bias, only if they statistically did something to improve fairness or bias. It is clear that this omission resulted in less discussion within the winners' reports on this important topic; what is included in the reports is predominantly captured in the section "Gender's role in improving forecasting accuracy." Our attempts to address this mistake are noted below.

NIJ hosted a virtual symposium in December 2021 to gather Challenge winners to discuss algorithm creation and fairness and bias involved in risk assessments. Racial bias and gender-specific needs were not directly related to the parameters of the Challenge but are crucial considerations for actuarial risk assessment creation and implementation. Therefore, Challenge Winners were asked to address the issues in panels and breakout sessions.

Summit discussions around these topics were recorded, transcribed, and then thematically coded by a team of research assistants and graduate research fellows. Participants highlighted barriers around addressing fairness, like the complexity of quantifying bias and competing definitions of fairness. Summit participants discussed the limitations of reducing bias through variable exclusion, suggesting that the dataset itself may contain upstream biases from earlier criminal legal processes.

Additionally, participants called for increasing the diversity of variables in datasets to better address female-specific needs, noticing that the dataset provided for the Challenge was lacking in this area. These topics are discussed in more depth in a recently published online article.[5] Themes of fairness in risk assessments are additionally expanded on with reference to extant literature in an academic research note.[6] The research note specifically references research that aligns with the points made by Summit participants, including the difficulty of eliminating bias due to competing definitions of fairness.[7]

Scholars recognize up to six definitions of fairness, making it a complex task to address.[8] Often, fairness boils down to balancing error rates or accuracy. Regarding gender-responsive needs, new figures indicate an increasing population of women involved in the criminal justice system.[9] This points to a growing need for gender-specific approaches to risk assessment.

In summary, research and the Summit participants highlight key considerations around fairness and gender-responsive needs in criminal justice actuarial risk assessment use. These discussions are crucial for improving risk assessments and corrections practices and programming oriented around risk assessments.

---

[5] Raven A. Lewis, Rachael M. Rief, and D. Michael Applegarth, "Best Practices for Improving the Use of Criminal Justice Risk Assessments: Insights from NIJ's Recidivism Forecasting Challenge Winners Symposium," Washington, DC: National Institute of Justice, January, 11, 2024, https://nij.ojp.gov/topics/articles/best-practices-improving-use-criminal-justice-risk-assessments.

[6] Rachael M. Rief, Raven A. Lewis, and D. Michael Applegarth "In Pursuits of Fairness: A Research Note on Gender Responsivity and Racial Bias in Criminal Justice Actuarial Risk Assessment," (forthcoming).

[7] See Jennifer L. Skeem and Christopher T. Lowenkamp, "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact," *Criminology* 54 no. 4 (2016): 680-712, https://doi.org/10.1111/1745-9125.12123; and Jennifer Skeem and Christopher Lowenkamp, "Using Algorithms To Address Trade-Offs Inherent in Predicting Recidivism," *Behavioral Sciences and the Law* 38 no. 3 (2020): 259-278, https://doi.org/10.1002/bsl.2465.

[8] See Richard Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 50 no. 1 (2021): 3-44, https://doi.org/10.1177/0049124118782533; and Kelly Roberts Freeman, Cathy Hu, and Jesse Jannetta, *Racial Equity and Criminal Justice Risk Assessment*, Urban Institute, https://www.urban.org/research/publication/racial-equity-and-criminal-justice-risk-assessment.

[9] Aleks Kajstura and Wendy Sawyer, Women's Mass Incarceration: *The Whole Pie 2024*, Northampton, MA: Prison Policy Initiative, March 5, 2024, https://www.prisonpolicy.org/reports/pie2024women.html.

Summary of the 2021 Challenge Winners' Reports

Twenty-six final reports from winning contestants, including one student report, were accepted and posted to NIJ's website. This paper focuses on the non-student reports (i.e., 25 reports). Additionally, not all analyses and discussions in this paper will be based on the remaining 25 reports due to inconsistent information provided in some of the sections of the reports, different forecast outcomes (recidivism in years 1, 2, or 3), or the type of models used.

To conduct this summary, every paper was read, and the following information was recorded: modeling strategies, data cleaning and feature engineering procedures (transforming data for the specific model to be used), and reported important variables. This information was systematically cataloged into a dataset with existing information like the categories each team placed in, winning Brier scores, and predicted probabilities (see the extracted data and dashboard for individual data for all the reports synthesized). More than one individual reviewed all papers to ensure inter-rater reliability.

A Better Score May Not Practically Matter

It should be noted at the outset of this paper that NIJ treats each of the winning reports equally. When comparing the winning submission results there are no practical difference in scores. While not noted in any of the winning papers, it was discussed at the Winners Symposium that the difference in the winning scores were often at the fifth or sixth decimal place of error.[10] And while that was important in deciding the winners, it does not lead to a practical difference in how well the algorithms performed. (Please see appendix for a thorough explanation of the impact of small differences in recidivism forecasts on the overall accuracy of the model).

Because of this, for the review that follows, we discuss all the winning team's modeling strategies and variables that were reported as equally important. The information about the models and the important variables is meant to yield an understanding of how teams created winning models and how their strategies apply to those working in the field in terms of data collection and analysis.

Additional Datasets Used

Seven teams used additional data. Among these teams, the most common was census data at the Public Use Microdata Areas (PUMA) level. Two teams reported additional outside data: CrescentStar reported accessing the Georgia Data Initiative[11] and Sevigny reported accessing the Environmental Protection Agency's (EPA) Environmental Justice Screening and Mapping Tool[12] and Traffic Proximity and Volume index,[13] the Bureau of Transportation's Local Area Transportation Characteristics for Households Data (LATCH) dataset,[14] Housing and Urban Development's Affirmatively Furthering Fair Housing Data and Mapping Tool,[15] the U.S. Department of Agriculture's Food Access Research Atlas,[16] and the University of Michigan's National Neighborhood Data Archive.[17]

[10] D. Michael Applegarth, Raven Lewis, and Rachael Rief, "Imperfect Tools: A Research Note on Developing, Applying, and Increasing Understanding of Criminal Justice Risk Assessments," (in press).

[11] "Data Tables," University of Georgia: Carl Vinson Institute of Government, https://georgiadata.org/data/data-tables.

[12] EPA's Environmental Justice Screening and Mapping Tool (EPA EJScreen), Version 2.3, United States Environmental Protection Agency, https://ejscreen.epa.gov/mapper/.

[13] "EJScreen Map Descriptions," United States Environmental Protection Agency (EPA), https://www.epa.gov/ejscreen/ejscreen-map-descriptions#:~:text=TrafficProximityandVolume,DatabaseHighwayPerformanceMonitoringSystem.

[14] "Local Area Transportation Characteristics for Households Data," United States Bureau of Transportation, https://www.bts.gov/latch/latch-data.

[15] "Affirmatively Furthering Fair Housing Data and Mapping Tool," United States Department of Housing and Urban Development, https://egis.hud.gov/affht/.

[16] "Food Access Research Atlas," U.S. Department of Agriculture: Economic Research Service, https://www.ers.usda.gov/data-products/food-access-research-atlas/.

[17] "The National Neighborhood Data Archive," University of Michigan, https://nanda.isr.umich.edu/.

According to most contestants' reports, additional data seemed to increase the accuracy of their models, but the increase was small and would likely not aid practice, thereby implying that, when additional outside data was included, the added variables tended not to be statistically significant. This may be due to the granularity employed to reduce disclosure risks of individuals in the dataset. For example, the PUMA designation regions were quite large, and multiple were often PUMAs combined, likely impacting their predictive power.

Consideration of Race and Gender

While many teams were only concerned with the overall accuracy of their submissions for each year, some winning teams attempted to improve the accuracy of scores by considering specific modeling strategies for males and females. Four of the 24 teams discussed the creation of gender-specific final models (i.e., different models used to separately predict male and female recidivism). Team PASDA, for example, stated that they used different ensemble models for males and females. Other techniques (e.g., random forest models) may account for race and gender by creating decision trees based on race and gender (i.e., splitting the data by male or female, or black or white); however, no teams noted if this was the case.

Feature Engineering

Many of the winning teams described the feature engineering methods used in their reports. Feature engineering aims to transform data to make it useable for specific types of analyses and to introduce field expertise. Transformations may include, but are not limited to, re-coding or changing the initial format of variables into formats that are useable for a given modeling strategy, dealing with missing values, and constructing new variables out of the existing variables (see exhibit 1).

Exhibit 1. Types of Feature Engineering

Categorical Encoding	Missing Data Imputation	Feature Extraction
Examples: Text variables -> numeric; Dummy coding (0-1)	Examples: Replacing with mean, median, mode; Complete case analysis	Examples: Principal Component Analysis (PCA); Literature, expertise, or theory

Note: Three types of feature engineering are listed here with implementation examples.

For example, creating a new variable using categorical encoding can be done by grouping existing data (e.g., to create variables like "total arrests" or "total convictions"). An example using feature extraction could involve parsing out existing variables to create new ones (e.g., transforming a variable representing percentages into two variables, one representing the numerator and the other the denominator).

Teams' reports describe their data cleaning process, including how they re-coded variables to fit their modeling strategies. At least 12 of the winning teams reported imputing missing values for some of their variables. However, which variables were imputed differed between teams and, in some cases, between different models by the same team. Additionally, nine teams discussed their methods and process for creating new variables or features (see exhibit 2).

Exhibit 2 summarizes the methods of feature creation discussed in the reports. Teams that created additional variables did so based on analysis like PCA or prior research and theory. For example, three teams discussed their use of PCA during their data preparation. Three teams also added prior-year probabilities to predict later-year probabilities as features in their models. For example, they used year 1 probabilities to predict risk of recidivism in year 2 and year 1 and 2 probabilities to predict risk of recidivism in year 3. Only two teams created features based on prior research. Four teams also created summative variables.

Exhibit 2. Variable Creation Explanation Summary and Counts

Methods and Reasons for Creating New Variables	Teams (#)*
Summative variables (e.g., total arrests or total convictions)	4
Conducted PCA	3
Prior year probabilities	3
Variables used in prior research	2
Separated existing variables	1
Team's own theory or hypotheses	1
Interaction terms (account for interactive effects)	1

Note: Nine teams reported creating their own variables in their reports, and of these teams, six different reasons were offered to support the creation of a new variable, and in some cases a single team offered multiple reasons. The number of teams that reported creating a variable for each of the six reasons is also listed.

** Counts are based on the authors' interpretations of the winners' reports.*

Analytical Approaches of Contestants' Models

The winning teams reported using a variety of analytical approaches and software. The reported methods and software were sorted into four categories: decision trees, neural networks, regression, and classification (see exhibit 3; see also an NIJ machine learning primer for practitioners for a description of machine learning and what these specific methods are[18]). The most common category reported by the winning teams included a decision tree approach (17 out of 25), followed by neural networks (10 out of 25), regression (7 out of 25), and classification (3 out of 25). Additionally, 17 out of 25 teams reported using an ensemble approach (a combination of analytical approaches). Teams that used more than one method or software to develop their models were counted separately in each reported category.

Exhibit 3. Broad Categorization of the Methods and Software Winners Reported Using in the 2021 Recidivism Challenge

Category	Reported	Type	Number of Teams
Decision trees *(17 total teams)*	XGBoost	Method	9
	LightGBM	Software	3
	CatBoost	Method	4
	Extreme Boost	Method	3
	Random Forest	Method	4
Neural Networks *(10 total teams)*	Artificial Neural Network	Method	6
	MLP (multilayer perceptron)	Method	2
	Deep Neural Network	Method	1
	TabNet	Software	1
Regression *(7 total teams)*	Lasso	Method	4
	Logistic Regression/Binary Logit	Method	6
	Ridge Regression	Method	1
Classification *(3 total teams)*	Support Vector Machine	Method	2
	WEKA	Software	1

Note: We have broken the methods reported by the challenge winners into four broad machine learning approaches: decision trees, neural networks, regression, and classification. In this table, we list all of the methods or programs the winners mentioned, whether a library package, software, or method, and include the number of teams reporting each. Teams may have used more than one method or software to develop their models.

[18] NIJ anticipates this to be an ongoing site, updated with new content as machine learning is adopted in innovate ways within the criminal justice system.

Variable Analysis

Out of the 25 reports considered for this analysis, 23 reports contained sufficient information to compare and analyze important variables. We considered the importance of variables in winning models in order to identify which variables more precisely predicted one's risk of recidivation. This information may help reduce recidivism for individuals on parole when used to inform resources that could decrease their risk. It should be noted that actual variable importance in many of the winning models was hard to identify because of the statistical models used; this will be discussed in the section Important Variables in Contestants' Models.

Most contestants provided lists of variables that were ranked in order of importance to the model; some included measures of feature importance, such as standardized coefficients (i.e., a standardized measure of how much a variable influences or improves the prediction). A few contestants simply listed important variables in their model without providing any additional analytical information. We categorized the variables reported by contestants based on what we could infer that their model was predicting:

1. Year 1 Recidivism or predicting the probability an individual will recidivate during year 1.
2. Year 2 Recidivism or predicting the probability an individual will recidivate during year 2.
3. Year 3 Recidivism or predicting the probability an individual will recidivate during year 3.

Within each category, we counted the number of times a variable was in a contestant's top five most important variables.[19] We also recorded any mention of a variable being in the top five, including instances when it was not possible to attribute the variable to a specific year.

Important Variables in Contestants' Models

It was common for contestant models to report using tens of variables. Therefore, we limited our analysis of important variables to the top five variables mentioned across reports (see exhibit 4). We relied on contestants' self-reporting and analysis of variable importance. More information on the relative importance of each variable for a given model can be found in contestants' reports (see winners' reports). We note that even if a variable is not a top five variable, it may have still been included in the overall model.

[19] In some cases, when contestants provided separate important variables for different groups (e.g., males vs females, Black vs. white individuals) more than five variables for a single team were included in this analysis. In these cases, the variables by which the data was split (e.g., gender, race) were also included in the analysis.

Exhibit 4. Percent of Teams Reporting a Variable as Top Five in Their Prediction for Each Forecasted Year and Any Mention

		Year 1	Year 2	Year 3	Any Mention
Total Teams Reporting Variables (N)		10	8	11	23
Variables	Gang_Affiliated	80%	38%	55%	70%
	Age_at_Release	80%	63%	82%	65%
	Percent_Days_Employed	NA	88%	73%	61%
	Jobs_Per_Year	NA	88%	64%	61%
	Prior_Arrest_Episodes_Felony	70%	0%	18%	48%
	Avg_Days_per_DrugTest	NA	63%	55%	48%
	Feature Engineered*	30%	50%	55%	35%
	Gender**	20%	0%	27%	30%
	Prior_Arrest_Episodes_PPViolationCharges	40%	13%	18%	30%
	Supervision_Risk_Score_First	40%	13%	0%	26%
	Prison_Years	40%	0%	9%	26%
	Prior_Arrest_Episodes_Property	50%	0%	0%	26%
	Residence_PUMA	10%	13%	9%	9%
	DrugTests_THC_Positive	NA	25%	0%	9%
	Race**	0%	0%	9%	4%
	Supervision_Level_First	0%	0%	9%	4%
	Prison_Offense	10%	0%	0%	4%
	Prior_Arrest_Episodes_Misd	0%	0%	9%	4%
	Prior_Conviction_Episodes_Viol	10%	0%	0%	4%
	Prior_Revocations_Parole	10%	0%	0%	4%
	Prior_Revocations_Probation	0%	0%	0%	4%
	Condition_MH_SA	10%	0%	0%	4%
	Condition_Cog_Ed	0%	0%	0%	4%
	Violations_ElectronicMonitoring	NA	0%	0%	4%
	Violations_Instruction	NA	0%	9%	4%
	Delinquency_Reports	NA	0%	0%	4%
	Program_Attendances	NA	0%	9%	4%
	Residence_Changes	NA	0%	0%	4%

Note: Dark shaded cells represent a higher percentage of teams reporting that variable as important in that year; progressively lighter shades represent lower percentages of teams reporting that variable as being in their top five.

** Feature-engineered variables are defined and described more specifically below.*

*** These percentages include both teams that identified Race or Gender as an important variable, and those that provided separate important variables for men vs. women and/or white vs. Black individuals.*

When comparing year 1 winning models to year 2 and year 3 winning models, there is a noticeable change in which variables were found as top five most important. Year 1 variables were "Prior Arrest Episodes Felony" (indicated as important in 70% of year 1 models), "Gang Affiliated" (80%), and "Age at Release" (80%). It is important to note that these tend to be variables that are considered static (i.e., nothing correctional officers can do to change these).

Models for year 2 and year 3 were able to include more dynamic variables (e.g., frequency of drug testing, employment data, housing, and compliance with terms of probation). Dynamic variables are those that are changeable, such as substance abuse and negative peer association. Static variables are recognized as a part of individuals' histories that are predictive of recidivism but are not responsive to interventions like treatment programs.[20]

Despite the introduction of dynamic variables, static variables like age and, to an extent, gang membership remained important across all years. It is important to point out that other static variables were not important once dynamic variables were introduced. "Prior Arrest Episodes Felony" dropped to 0% and 18% as a top five predictor for years 2 and 3, respectively. To a lesser extent, there was also a reduction in the top five presence of gang affiliation variables. Gang Affiliation dropped to 38% of teams in year 2, and 55% of teams reported year 3 variables. The dynamic variables in year 2 and year 3 models that were found to be important focused on supervision activities like drug testing and percent days employed, and were thus of greater importance to models.

Many of the features engineered by contestants were consistently identified as being important to their models. The most common engineered feature across contestants' submissions was "total arrests," a composite variable comprising all arrests. Four teams (PASDA, Crescent Star, KMG_BQR, and CategOracles) engineered a variable for "total arrests." For one of these teams, "total arrests" was one of their top five predictive variables. Interestingly while "total arrests" was a top five predictive variable for this team, the variables for individual arrest categories (e.g., Prior Arrest Episodes Felony, Prior Arrest Episodes PPViolationCharges) appeared as less predictive in the model. Therefore, combining variables to create a new one decreases the ability to explain if it is just a part of the new combined variable that matters or all parts of it equally (i.e., all arrest variables were collapsed so they could not be differentiated). However, it may increase the predictive power of the model, in part because it doesn't have to be as specific in how it contributes to the model.

Three teams (DEAP, KMG_BQR, and Oracle) included predicted probabilities of recidivism from the prior years in their model to predict future probabilities of recidivism. For example, a team used year 1 probabilities to predict the probability of recidivating in year 2, and year 2 probabilities to predict the probability of recidivating in year 3. This feature, the prior year predicted probabilities, was important in these three teams' models. This suggests that knowing an individual's prior recidivism likelihood may help glean more information about someone's likelihood of recidivism in the future.

Two teams (CategOracles and Duddon Research) used categorical encoding to create a new variable that indicated if there is missing data in the "Jobs Per Year" variable. This new binary variable would equal one if an individual's data for "Jobs Per Year" was missing and zero otherwise. This new variable was associated with a decreased likelihood of recidivism.[21] Similarly, one team (Idle speculation) created separate variables that attempted to split the numerator and denominator of some of the provided variables. Both their estimated numerator of "percent days employed" and the denominator of "jobs per year" were identified as important in their Year 2 and Year 3 model forecasts. This suggests that accurately collecting data points, such as the number of days employed and the number of jobs one has started during supervision, may be more important in predicting future recidivism than the transformed "Jobs Per Year" and "Percent Days Employed" variables.

---

[20] Ruth E. Mann, R. Karl Hanson, and David Thornton, "Assessing Risk for Sexual Recidivism: Some Proposals on the Nature of Psychologically Meaningful Risk Factors," *Sexual Abuse: A Journal of Research and Treatment* 22 no. 2 (2010): 191-217, https://doi.org/10.1177/1079063210366039.

[21] This is separate from missing data imputation methods described above since it involves creating a new variable to indicate that data about a given variable was missing.

Discussion

No Single Best Modeling Strategy for Forecasting Recidivism

In terms of modeling strategies, teams used different approaches. Thus, we wondered how different modeling strategies inform results. An analysis showing the differences in Brier scores among winning models (see appendix), showed only a small difference in Brier scores between the worst-performing and the best-performing models, suggesting that different modeling strategies do not yield large practical differences in the ability to accurately predict recidivism. In fact, we note that the degree to which models are fine-tuned prior to calculating risk does not appear to matter in this Challenge — simpler models performed similarly to more complex, fine-tuned ones. This conclusion mirrors recent work that found that newer machine learning algorithms had higher predictive validity but the range between the performance of the best and worst algorithms remained small across 10 different scenarios.[22] Thus, while more research is needed in this area, we conclude that, for practitioners, it may be more realistic to focus on using simpler models with transparent results during the creation and implementation of risk assessment tools.

Improved Model Performance as a Common Goal

The Challenge contestants created their models with the goal of high performance, meaning they cared more about the predictive accuracy and less about the interpretability of the model.[23] In fact, many of the contestants gave up almost all their transparency and interpretability for small gains in performance. We concluded that there was no single superior model used by winners. In fact, the simpler models appeared to perform practically as well as the more complex ones. Moreover, whether models are fine-tuned prior to calculating risk did not appear to matter in this Challenge. Conversely, practitioners and, to an extent researchers, who create risk assessment tools need to be transparent so they can easily translate the results into practical goals and be responsive to community questions and concerns. Practitioners and modelers need to work together to determine the acceptable tradeoff between the ability to explain and accuracy.

Current practice could benefit from using more explainable machine learning models such as decision trees and regression models. These methods, paired with stakeholder-informed and data analysis-driven, feature-engineered variables, can result in fairly accurate models that are explainable. Additionally, we see examples where incorporating fairness into a model design can yield fairer results with minimal compromises to accuracy. This finding requires further investigation for more generalizability but is in line with other research that looks at the tradeoff of machine learning fairness vs. accuracy.[24,25]

The Role of Gender in Improving Forecasting Accuracy

Winners appeared to better predict female recidivism despite the majority use of gender-neutral modeling strategies. Moreover, for teams that considered the importance of variables, gender was among the list of top five variables for many teams (see exhibit 3). This aligns with the idea that gender-specific risk assessments that include gender-responsive risk factors are superior to gender-neutral ones in the forecast of female recidivism. We note that a few winning participants

---

[22] Grant Duwe and KiDeuk Kim, "Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms To Predict Recidivism," *Criminal Justice Policy Review*, 28 no. 6, (2017): 570-600, https://doi.org/10.1177/0887403415604899.

[23] Some participants ran separate models for males and females to improve their predictions and test for gender differences. We did not draw conclusions about whether this helped the predictive accuracy of models because few teams indicated that they used this strategy. More research is needed to determine if this modeling strategy is effective in predicting recidivism by gender.

[24] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani, "Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy," *Natural Machine Intelligence* 3, (2021): 896-904, https://doi.org/10.1038/s42256-021-00396-x.

[25] Emily Black, Manish Raghavan, and Solon Barocas. *Model Multiplicity: Opportunities, Concerns, and Solutions. In FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY: Association for Computing Machinery, (2022): 850-863, https://doi.org/10.1145/3531146.3533149.

discussed separate modeling strategies for each gender. However, since so few teams tested this strategy of gender-specific models, we cannot draw meaningful conclusions. Further research is necessary to determine if this modeling strategy is effective in predicting recidivism by gender. Themes on the role of gender in risk assessment were further discussed at the symposium in a panel on gender-specific needs.[26]

Models' Ability To Identify Important Variables Should Be Considered

Based on the winners' reports, it is not always possible to identify whether a variable increases or decreases their model's likelihood of forecasting recidivism. For some machine learning methods like neural networks, it is not currently possible to know what variables contribute to the model. When using these more opaque methods, such as neural networks, teams instead relied on other analyses to identify variable importance — for example, measuring the positive or negative correlation between variables and the recidivism outcome.

Additionally, while a variable may be in the top five most important in a model that accurately forecasts recidivism risk, it is not always possible to determine the exact nature of the relationship between that variable and risk to recidivate. Machine learning methods challenge research-minded practitioners to not only know whether something is predictive but to also understand why it is predictive. As a result, machine learning methods will result in better risk assessment tools and programming that serves case managers and the individuals under supervision. However, they will also require more insights from the practitioners to respond to the results in a meaningful and impactful way.

Important Variables

In addition to the type of model, we looked at the variables used in teams' models. Every team included slightly different variables, with some engineering their own features. As described in their reports, it was common for the team's final models to include all provided and feature-engineered variables. This likely increased the accuracy of their model. However, this approach may have high multi-collinearity (when two or more independent variables are correlated with each other, making it difficult to determine the effect of each variable independently), which may lead to misleading and complicated results. Multi-collinearity may disguise or at least make it more difficult to determine the importance of variables. Practically, this suggests that recording information on all possible variables is not necessarily helpful or useful. This approach could be unnecessarily expensive, since the resources needed to collect and store additional data may have diminishing returns on increased model accuracy.

Therefore, the reports from the winning contestants offer important insights into modeling strategies that may be useful in the creation and application of risk assessment tools. For agencies creating, adopting, and implementing new tools, design considerations are critical. The details from the reports summarized here, paired with the design and implementation considerations mentioned by winning teams in a recent symposium[27] can aid agencies in the creation and implementation of their own risk assessment tools.

Variables Predictive of Recidivism Change With Time Since Release

The variables predictive of recidivism are inconsistent across years of supervision. This suggests that the variables most accurately correlated with the actual outcome of recidivism may change the longer an individual is under supervision. To address these issues, case managers should assess an individual's risk of recidivism throughout their reentry experience, not just at the beginning of reentry. More regular risk assessments could also aid in better allocation of supervision resources and adaptation of risk assessment tools to incorporate local supervision metrics. This could allow for case managers to better communicate current risk factors to individuals.

[26] Rachael Rief, D. Michael Applegarth, and Raven Lewis, "In Pursuit of Fairness: A Research Note on Gender Responsivity and Racial Bias in Actuarial Risk Assessment," (in press).

[27] Applegarth, Lewis, and Rief, "Imperfect Tools."

Although teams used different methods, the most consistent top variables among teams were age at release, indicators of gang affiliation, percent days employed, and the number of jobs per year. These patterns, while not causal, lend themselves to important practical considerations for case managers responsible for carrying out or interpreting risk assessments that predict recidivism. They also highlight an area for future research. Specifically, additional research is necessary to determine whether revaluation after the first year to include the newly available information would be feasible.

Static and Dynamic Variables Can Both Be Predictive of Recidivism

As previously mentioned, variables predictive of recidivism changed with time since release, with important variables predicting recidivism in year 1, like prior felony arrest, being less important in years 2 and 3. Except for age, which mattered across all years, these results suggest dynamic variables that an individual can change with time may be more useful in predicting recidivism than static variables that an individual cannot change. A future direction may be to create new variables that are more dynamic in nature (for example, seeing how the static variable, prior felony arrest, compares to a more dynamic counterpart, such as employment status during parole, in predicting recidivism).

On a related note, multiple teams engineered their own features, like total arrests. This variable, specifically, increased the predictive accuracy of teams' models, but also reduced the ability to explain what this variable means because information was lost in creating one variable representing all arrests. Additionally, if someone is identified to be higher risk due to dynamic factors, case managers could, if resources are available, direct individuals to evidence-based programs and services to address their specific dynamic risk.[28] Therefore, the Challenge results suggest these static variables may be useful for forecasting recidivism in the first year following release, whereas more dynamic variables are more predictive of long-term recidivism and allow for case managers to create individual responses to individuals under their supervision.

Future Research for Proxy Supervision Variables Upon Release or While in the Facility

Future research could incorporate more dynamic variables at the time of release and see if they would perform better than the static variables identified here as important. Given the added improvement in forecast accuracy produced through some of the supervision variables (e.g., "percent days employed", "jobs per year"), it may be beneficial to identify proxy measures for these variables while an individual is incarcerated, or earlier following release, to aid predictive models. For example, employability measures such as "job readiness" or "job prospects," while an individual is incarcerated may be examples of predictive proxy measures to track and study in the future. This could potentially lead to more accurate first year forecasts and, more importantly, better targeted employment programming services prior to someone's release. Moreover, employability is, as stated above, mostly dynamic. Employability could be assessed to inform targeted services ranging from general education, specialized training certifications, and behavior skills training[29] for specific needs. If such programming was provided before reentry, then an individuals' success in these programs could be a proxy measure of employability.

[28] August F. Holtyn et al., "Employment Outcomes of Substance Use Disorder Patients Enrolled in a Therapeutic Workplace Intervention for Drug Abstinence and Employment," *Journal of Substance Abuse Treatment* 120, 108160 (2021), https://doi.org/10.1016/j.jsat.2020.108160.

[29] Elizabeth Lin et al., "Behavioral Skills Training for Teaching Safety Skills to Mental Health Service Providers Compared to Training-as-Usual: A Pragmatic Randomized Control Trial" *BMC Health Services Research* 24, 639 (2024), https://doi.org/10.1186/s12913-024-10994-1; and Kristen M. Brogan, John T. Rapp, Odessa Luna, and Steven P. Lafraniere, "The Role of Applied Behavior Analysis in Juvenile Justice Settings," *Corrections Today* 82 no. 3 (2020): 22-27.

Engineered Features Were Important in Challenge Contestant Models

As machine learning becomes more common in risk assessments, practitioners should consider how to include their field expertise. Feature engineering, specifically feature extraction, may be a particularly tractable approach to integrating this outside information. For example, Challenge teams described the creation of new variables or features based on past research. One team created a variable indicating whether a female on parole had dependents because research suggests that women with children younger than 18 are less likely to commit a subsequent offense after release.[30,31]

Limitations in Interpreting Challenge Winner Reports

There were limitations to our analysis of finalist reports and the Challenge itself. First, despite all participants being provided with the same prompt questions and topics to include in their reports, there was a high degree of variability in the winner's reports' background, depth, and detail. Secondly, many of the Challenge participants noted some issues with the data, some of which were intentional, stemming from the need to protect the identities of individuals on parole and others unknown before the Challenge. Finally, the analysis of winners' reports and findings highlighted some new questions and future challenge design considerations for the Challenge design team.

There were a few limitations of the Challenge dataset. First, the dataset did not include complete information for all variables, like gang membership. For example, women were omitted from the gang membership variable because very few women in the sample fell into this category. Some variables in the dataset may have also been indirectly associated with the outcome due to data entry patterns. The potential "leakiness" could have inflated the importance of the applicable variables, meaning that some variables important in the Challenge might not be important in practice. Additionally, it is unknown if the variables in the Challenge dataset would perform differently with more precise data, as many data points were aggregated to reduce the risk of disclosure for individuals in the dataset (see Challenge Dataset description). As such, it is important to emphasize that these are the significant variables specific to the Challenge and should not be assumed to be significant in general. Community corrections departments are encouraged to go through a similar process with their own data at the level of granularity that is available.

There were also a few limitations to the Challenge structure. Since the Challenge focused on model performance and not being able to identify important variables, most if not all entries did not try to eliminate multi-collinearity (the extent to which two or more variables explain or predict the same "parts" of recidivism). Testing for multi-collinearity could help to better identify unique variables, instead of redundant ones. When there is multi-collinearity, usually one of three things happens. First, one of the variables may be omitted (this is the least likely), which leads to more model error. Second, a single variable may be created out of the variables that have multi-collinearity with each other. This affects how the results are interpreted for this new variable and potentially produces more model error. Third, all the variables are left in the model. This typically improves (non-adjusted) model performance, but it can mask variable importance. Variable importance is usually measured as the amount of unique contribution a variable has to prediction; by leaving all the variables in, variables with multi-collinearity will have less unique contribution. Practically, this would lead to missing important variables that affect recidivism and that researchers and practitioners could put their resources toward. There was also no incentive for entries to limit the number of variables that were included in the final model, as prizes were all based on model error.

---

[30] Brent B. Benda, "Gender Differences in Life-Course Theory of Recidivism: A Survival Analysis," *International Journal of Offender Therapy and Comparative Criminology* 49 no. 3 (2005): 325-342, https://doi.org/10.1177/0306624x04271194.

[31] Nancy J. Harm and Susan D. Phillips, "You Can't Go Home Again: Women and Criminal Recidivism," *Journal of Offender Rehabilitation* 32 no. 3 (2001): 3-21, https://psycnet.apa.org/doi/10.1300/J076v32n03_02.

When creating a risk assessment tool, forecast models with fewer variables may be easier to interpret and to communicate forecasted risk to case managers and the individuals under their supervision. Simpler risk assessments (those with fewer variables to track and analyze) may also be easier to repeatedly use in practice. While the Challenge showed that simpler models performed just as well, we were unable to explore if "simpler" in terms of the number of variables used also perform practically as well. Future creators of Challenges should consider incentivizing model simplicity and interpretability in addition to accuracy and fairness.

Conclusion

This paper aims to add to the knowledge of risk assessment creation by summarizing the Challenge winners' papers. We focused on the methods described in 25 non-student reports. After systematically recording each of the papers' methods, we summarized important details like their models, their variables, whether they used additional data, and whether they split the data by race and gender. NIJ hopes that, although the Challenge has concluded, the summary of the winners' papers continues a meaningful conversation of risk assessment creation. The summary of winners' reports suggests that as research on risk assessment continues, there should be an increased focus on design considerations. Additionally, this Challenge focused on fairness and accuracy through narrow definitions, where several other options exist.[32]

We concluded that there was no single superior model used by winners. In fact, the simple models appeared to perform practically as well as the more complex ones. Moreover, whether models are fine-tuned prior to calculating risk did not appear to matter in this Challenge. Some participants also discussed separate modeling strategies for males and females, but we did not draw conclusions about whether this helped the predictive accuracy of models because so few teams indicated that they used this strategy. Thus, more research is needed to determine if this modeling strategy is effective in predicting recidivism by gender.

Current practice could benefit from using more explainable machine learning models such as decision trees and regression models. These methods, paired with stakeholder-informed and data-analysis-driven feature-engineered variables, can result in fairly accurate models that are explainable. Additionally, we see examples where incorporating fairness into a model design can yield fairer results with minimal compromises to accuracy. This finding requires further investigation for more generalizability but is in line with other research that looks at the tradeoff of machine learning fairness vs. accuracy.[33,34]

In addition to the type of model, we looked at the variables used in teams' models. Every team included slightly different variables, some engineering their own features. When digging deeper into the importance of the variables, it appeared that variables predictive of recidivism were different across each year an individual is on supervision. In other words, variables predictive of recidivism changed with time since release, with important variables predicting recidivism in year 1, like prior felony arrest, being less important in years 2 and 3. Except for age, which mattered across all years, these results suggest dynamic variables that an individual can change with time may be more useful in predicting recidivism than static variables that an individual cannot change. A future direction may be to create new variables that are more dynamic in nature (for example, seeing how the static variable, prior felony arrest, compares to a more dynamic counterpart, such as employment status during parole, in predicting recidivism). Additionally, multiple teams engineered their own features, like total arrests. This variable, specifically, increased the predictive accuracy of teams' models, but also reduced the ability to explain what this variable means, because information was lost in creating one variable representing all arrests.

We argue the Challenge met its aim of a better understanding of factors contributing to recidivism. In addition, we also gained a better understanding of the limitations of the methods, models, and feature creation used to forecast recidivism. Altogether, this review of winner's reports suggests a need to continue evaluating challenging elements of machine learning models.

---

[32] See Berk et al, "Fairness in Criminal Justice Risk Assessments."

[33] Kit T. Rodolfa, Hemank Lamba, and Rayid Ghani, "Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy," *Nature Machine Intelligence* 3 (2021): 896-904, https://doi.org/10.1038/s42256-021-00396-x.

[34] Emily Black, Manish Raghavan, and Solon Barocas, "Model Multiplicity: Opportunities, Concerns, and Solutions." In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, 850–863, https://doi.org/10.1145/3531146.3533149.

The summary of these winning papers provides important information related to the creation of risk assessments. These include considerations of the types of models used and whether there is a superior model. Winning reports also brought up points related to variables created and used and that, ultimately, were important in their models. NIJ hopes that although the Challenge has concluded, the summary of the winners' papers continues a meaningful conversation of risk assessment creation.

Appendix

Why Winning Teams Had Practically Equivalent Results

We also conducted an analysis looking at winning teams' non-winning submissions; that is, if a team won a one-year male Brier score prize, we also looked at the scores of their other submissions that did not win a prize. These additional non-winning scores are practically indistinguishable in Brier scores compared to the winning Brier scores. This finding implies that the various modeling approaches by winning contestants are practically, not statically, just as predictive as the most accurate winning model. Therefore, this allows us to combine all winning model papers. Additionally, fairness and ease of application to the field can potentially be incorporated into the selection of a model without a drastic reduction in accuracy.

Predicted probabilities were between 0% and 100%. Actual recidivism was binary, with 0 meaning the individual did not recidivate and 1 meaning they did recidivate. The predicted probabilities and the actual results were compared in order to estimate the error of submitted models. Specifically, a single Brier score for each challenge submission was calculated to estimate a model's error, as described in more detail on the challenge website and in Results from the National Institute of Justice Recidivism Forecasting Challenge. Therefore, a Brier score is between 0 and 1, where the closer a score is to 0, the better (i.e., more accurately) the model performed.

A Brier score is the average of the squared error. To demonstrate how a Brier score estimates model error, we looked at eight specific examples (see exhibit 5). In the case where an individual was given a 0% prediction of recidivating, but they actually did recidivate that year, then the squared error for that individual would be 1 (i.e., the absolute value of the actual result minus the predicted probability squared = (| 1 – 0|)^2 = 1). Exhibit 5 shows eight examples of calculating the squared error used in an overall Brier score. We demonstrate four example prediction probabilities (i.e., 0%, 25%, 50%, and 100%) with two actual realizations (i.e., Recidivated and Did Not Recidivate).

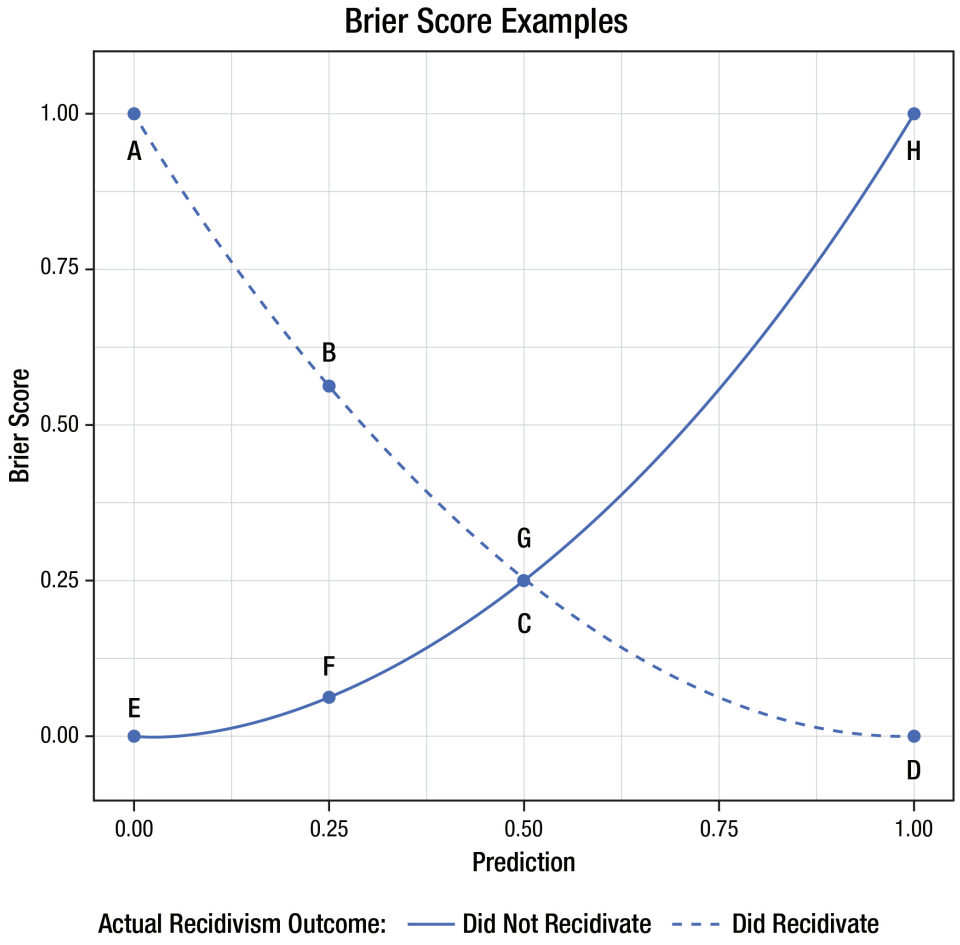
Exhibit 5. Example of Brier Score Outcomes

Brier Score Based on Individual Prediction and Actual Result		Predicted Probability			
		0% Prediction	25% Prediction	50% Prediction	100% Prediction
Actual result	Recidivated (1)	1[A]	|1-.25|^2 = .5625[B]	|1-.5|^2 = .25[C]	0[D]
	Did Not Recidivate (0)	0[E]	|0-.25|^2 =.0625[F]	|0-.5|^2 = .25[G]	1[H]

Note: This exhibit shows eight examples of Brier scores, a prediction probability, and whether the individual recidivated. Four example prediction probabilities are demonstrated (i.e., 0%, 25%, 50%, and 100%) with two actual realizations (i.e., Recidivated and Did Not Recidivate). These scores can directly map the points on exhibit 6. If, instead, we wanted to calculate a single Brier score based on the eight examples, we would obtain a Brier score of 0.3906 (i.e., [1 + 0.5625 + 0.25 + 0 + 0 + 0.0625 + 0.25 + 1] /8 = 0.3906).

Exhibit 6 shows the full range of Brier scores based on a single prediction probability between 0% and 100% and whether the individual recidivated. In the Challenge, a submission's Brier score was based on the average of these squared errors across the predicted probabilities for all individuals in the dataset. For example, instead of calculating Brier scores for each predicted probability in exhibit 5, we take the average of the squared error to calculate a Brier score. In this case, if we considered only the eight predictions below, we would obtain a Brier score of 0.3906 (i.e., [1 + 0.5625 + 0.25 + 0 + 0 + 0.0625 + 0.25 + 1] /8 = 0.3906).

Exhibit 6. Full Range of Possible Brier Scores for a Single Individual Based on Predicted Probabilities and Recidivism Outcome



Note: This graph shows the full range of Brier scores based on single-prediction probabilities between 0% and 100% and whether the individual recidivated. The points on the graph directly relate to the numbers shown in exhibit 5.

Though a lower Brier score may have resulted in more prize money for contestants participating in the Challenge, improvements to Brier scores among winning models may or may not be relevant to practitioners looking for a better risk assessment tool. We now explore how a small improvement in an algorithm's accuracy (i.e., Brier score) may not translate to any meaningful difference in practice.

Exhibit 7 compares the lowest winning Challenge Brier score and the highest winning Brier score in the female and male categories over the three years using a Brier Skill score. A Brier Skill score compares the model accuracy of two models. It is calculated by first computing the percent accuracy improvement by dividing a Brier score by the Brier score of the comparison model. The Brier Skill score is then 1 minus the percent accuracy improvement (see The NIJ Recidivism Forecasting Challenge: Contextualizing Results). As shown in exhibit 7, female Brier scores among winners were, at most, 1.21%, 5.3%, and 3.64% more accurate (Brier Skill score) than the lowest winning score in years 1, 2, and 3, respectively. Similarly, the Male Brier scores among winners were, at most, 1.14%, 8.49%, and 4.63% more accurate than the lowest winning score. However, there is no threshold in the literature as to what percent improvement is "significant" or not when interpreting a Brier Skill score. It is, therefore, up to researchers and practitioners to decide what is a meaningful improvement.

Exhibit 7. Accuracy Improvement Between Highest and Lowest Winning Challenge Brier Score via Brier Skill Score

	Brier Skill Score (Lowest Brier Score)		
	Year 1	**Year 2**	**Year 3**
Female Brier score	1.21 % (0.1719)	5.3% (0.1196)	3.64% (0.1139)
Male Brier score	1.14% (0.1900)	8.49% (0.1542)	4.63% (0.1463)

Note: This table shows the absolute percent improvement in Brier score between the highest and lowest Brier score winners in each of the male- and female-specific Brier score categories in years 1, 2, and 3 separately. Absolute percent improvement between two Brier scores is also known as a Brier Skill score.

As exhibit 6 shows, the change in the Brier score from an incorrect prediction is not linear; therefore, changing a prediction by X is also not linear. If we wanted to understand the effect of changing a single prediction by 0.25, we would have to know the original prediction. For example, if the prediction for someone that did not recidivate was 0.25, this originally would have had a Brier score of 0.0625, and if we changed the prediction from 0.25 to 0.50, it would have had a Brier score of 0.25, a difference of 0.1875. If the original predication had been 0.50 with a Brier score of 0.25 and we changed it to 0.75, we would now have a Brier score of 0.5625, a difference of 0.3125. This makes it very difficult to contextualize or put into practical terms what we would need to see in the data for these score differences. These examples are for a single prediction when the Brier score is the average of all the predictions.

Exhibit 8 does provide a toy example of how the difference in Brier Skill scores could have occurred. It summarizes the total number of individuals for the year 2 dataset. We will only consider year 2 since it had the largest accuracy improvement, as shown in exhibit 7. Exhibit 8 also outlines different toy examples (interpretable scenarios) in which the accuracy improvements in exhibit 7 could result.

Exhibit 8. Meaning of a 5.3% or an 8.49% Brier Skill Score

	Year 2	
	Female	**Male**
Brier Skill Score (i.e., Percent Accuracy Improvement)	5.3%	8.49%
Number of Individuals in the Dataset	742	4718
Example scenarios that could attain the above Brier Skill scores, assuming Model 1 had perfect accuracy		
Scenario 1: Number of individuals in Model 2 that would have a prediction probability of 50%	20.08 (2.71%)	269.47 (5.71%)
Scenario 2: Number of individuals in Model 2 that would have a prediction probability of 75% if they did recidivate or 25% if they did not recidivate	80.31 (10.82%)	1077.87 (22.85%)

Note: Perfect accuracy is where a model gave prediction probabilities of 100% for all individuals who recidivate in year 2 and 0% for all individuals who did not recidivate in year 2.

For the Challenge, a prediction probability over 50% was interpreted as forecasting an individual would recidivate in a given year. Therefore, a Brier score being off by a quarter may or may not have resulted in false negatives or positives in predicting recidivism, whereas a 0.5 difference in a prediction probability would have resulted in a difference in the interpretation (i.e., predicted to recidivate or not).

Based on the analysis provided in exhibit 8, there is a small enough difference between model accuracy among winning submissions to compare them directly. The 5.3% improvement for females could be interpreted as, in the worst case, one model incorrectly forecasting an additional 20 individuals in the dataset to recidivate when they did not. Alternatively, this could be interpreted as 80 individuals' prediction probability to be off by 25% (e.g., Model 2 predicting a 75% probability of recidivating compared to Model 1 predicting a 100% probability of recidivating for the same 80 people). For the Male Brier scores in year 2, this could be interpreted as, in the worst case, one model incorrectly forecasting an additional 269 individuals.

We have described our analysis to justify how the small differences in Brier scores between winners result in differences between models small enough to not be of practical relevance to practitioners. The fairness and accuracy metric is an extension of the Brier score calculated through a penalization of inaccurate predictions based on race. However, we did not look at fairness and its impact. This is work that NIJ intends to explore more deeply in a future publication.