UNIVERSITY of PENNSYLVANIA

Department of Criminology

Working Paper No. 2016-4.0

An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism

Richard Berk

An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*

Richard Berk
Department of Criminology
Department of Statistics
University of Pennsylvania

7/24/2016

*The entire project would have been impossible without the efforts of Jim Alibrio, Fred Klunk and their many colleagues working at the Pennsylvania Board of Probation and Parole and the Pennsylvania Department of Corrections. Thanks also go to the National Institute of Justice for financial support and to Patrick Clark who was the project monitor at NIJ.

Abstract

Objectives: The Pennsylvania Board of Probation and Parole has begun using machine learning forecasts to help inform parole release decisions. In this paper, we evaluate the impact of the forecasts on those decisions and subsequent recidivism.

Methods: A close approximation to a natural, randomized experiment is used to evaluate the impact of the forecasts on parole release decisions. A generalized regression discontinuity design is used to evaluate the impact of the forecasts on recidivism.

Results: The forecasts apparently had no effect on the overall parole release rate, but did appear to alter the mix of inmates released. Important distinctions were made between offenders forecasted to be re-arrested for nonviolent crime and offenders forecasted to be re-arrested for violent crime. The balance of evidence indicates that the forecasts led to reductions in re-arrests for both nonviolent and violent crimes.

Conclusions: Risk assessments based on machine learning forecasts can improve parole release decisions, especially when distinctions are made between re-arrests for violent and nonviolent crime.

Keywords: — Parole, Machine Learning, Recidivism, Forecasting, Regression Discontinuity Design

1 Introduction

Risk assessments have been used to inform parole decisions in the United States since the 1920s (Burgess, 1928; Borden, 1928). By and large, this practice has been treated as routine and sensible, even as the risk procedures became more heavily determined by actuarial methods (Pew, 2011). But, the past few years has seen increased scrutiny and controversy surrounding the use of actuarial risk assessments for a variety of criminal justice decisions (Gottfredson and Moriarty, 2006; Berk, 2012; Tonrey, 2014; Monahan and Skeem, 2015). The concerns are largely ethical (Harcourt, 2008) and jurisprudential (Starr, 2015). Still, there also seems to be some common ground (Berk and Hyatt, 2015).

A somewhat surprising oversight is that with very few exceptions (Berk et al., 2010b; Holsinger, 2013; Miller, 2013, McCafferty, 2015), the debates have unfolded with scant information about how actuarial risks assessments have affected practices and outcomes. Commonly, changes in practices and outcomes are ignored or assumed. In this paper, we evaluate the impact of machine learning forecasts of "future dangerousness" made available to the Pennsylvania Board of Probation and Parole.

We do not consider here whether in principle these forecasts were sufficiently accurate to improve practice. They were. 58% of the individuals released on parole after their first, mandatory hearing were arrested for a new offense within 2 years. If the Board only released individuals forecasted to not be re-arrested, test data projections determined that about 27% would have been arrested for a new offense within 2 years. We focus on is how the forecasts may have affected parole board decisions that, in turn, could affect parole outcomes after release.

2 The Setting

On September 29, 2008, the Governor of Pennsylvania, Edward Rendell, called for an independent review of the state's Department of Corrections and Board of Probation and Parole following the shooting deaths of two police officers by two paroled individuals. The review was conducted by Professor John Goldkamp from the Department of Criminal Justice, Temple University. Professor Goldkamp's initial review of the parole system identified a typology of offenders likely to commit a violent crime after release in parole.

In January of 2009, following two additional brutal murders by an individual on parole, the Governor called on the General Assembly to establish flat determinate sentences for repeat violent offenders and require a 5-year period of post-release supervision by the Board. A "repeat violent offender" designation would apply to anyone convicted as an adult or juvenile of two or more crimes of violence, or one crime of violence and a violation of Pennsylvania's Uniform Firearms Act. These measures were in addition to the steps already being taken by the Department of Corrections and the Board to better identify offenders with a history of violence, either in prison or on parole.

The exclusive focus on the incarceration of violent offenders may seem somewhat at odds with current concerns about the overuse of incarceration, but public safety was, and remains, one key objective of criminal justice practice. Subsequently, funding was provided by the National Institute of Justice to develop state-of-the-art risk assessment tools to help the parole board make better release decisions. "Big data" would be made available, analyzed with machine learning procedures.

Beginning in 2010, training data were provided by the Department of Corrections. Several machine learning procedures were applied. Random forests (Breiman, 2001) proved to be the most effective. Full development of the proposed forecasting procedures took several iterations as new data were made available and as the Board provided feedback on early results. A challenging step was linking the forecasting procedure to the available electronic data so that forecasts could be obtained as needed in real time. Another challenge was to make the forecasts available to parole board members in a fully accurate and easily accessible form. Eventually, random forests forecasts could on a routine basis be provided to the Board in a format that was easy to understand.

The development process was completed in the spring of 2013 after which a lengthy demonstration exercise began. Board members used the forecasts as parole decisions were considered, and data were collected on the decisions made and how the released individuals fared in parole. An evaluation of the demonstration exercise is the focus of this paper.

Three parole outcomes were to be forecasted: (1) an arrest for a crime defined as violent, (2) an arrest for a crime not defined as violent, and (3) no arrest. Included as violent crimes were murder, non-negligent manslaughter, simple assault, aggravated assault, robbery, arson, and rape. Arrests while under parole supervision and arrests whether under supervision or not were to be considered separately.

In addition to the forecasts, a measure of forecasts reliability was provided. As its name suggests, the random forest algorithm introduces some randomness into its forecasting machinery. There are sound statistical reasons for this approach, and one benefit is that a measure forecast reliability is available. In this setting, "reliability" means the degree to which the algorithm itself can consistently make the same forecast for a given inmate despite some random variation built into the random forests algorithm.

The evaluation was to address three related questions.

1. Did the overall proportions of inmates released by the Board change because of the forecasts?

2. Did the forecasts lead to changes in the kinds of inmates the Board released on parole?

3. What impact, if any, did the forecasts have on arrests after an individuals were paroled?

All three questions were addressed for four different classes of inmates whose circumstances with respect to a parole hearing could be rather different. In this paper, we focus exclusively on inmates who were being considered for parole for the first time after having served their minimum required sentence. The other three classes of inmates were making return visits to the parole board because they were initially turned down, or after being paroled, were returned returned to prison. In board brush strokes, the findings reported here are the same for all four groups, but role of the risk forecasts is necessarily much more complicated for inmates who are repeating the parole hearing process. A proper discussion of how these inmates fared is well beyond page constraints of this paper. The full report to the Board with that material included can be provided upon request.

3 The Impact of The Forecasts on Board Decisions

The purpose of the forecasts was to provide new information to the board members that would help them better assess the risks a prospective parolee posed. One of three forecasts was provided for each case: an arrest for a violent crime, an arrest for a crime that was not violent, and no arrest.

In addition, the random forest procedure internally calculates its performance reliability, which provides information about the reliability of all forecasts. A reliability can range from 0.0 to 1.0, with 0.0 meaning unreliable and 1.0 meaning perfectly reliable. For ease of use by the Board, the reliability values were organized into three levels: low, medium or high reliability. A value less .4 was considered low, a value greater than .5 was considered high, and a value between .4 and .5 was considered moderate.[1]

The key decision made by the Board is whether to release an inmate on parole. Consequently, an important outcome studied was whether a decision was made to grant parole. We anticipated that having available a forecast of a new arrest, especially for a violent crime, would reduce the likelihood of a parole release, but only if the forecast was sufficiently reliable. Unless the forecast had sufficiently reliability, it would be effectively ignored. An initial goal of the evaluation was to determine whether these expectations were borne out.

A second evaluation question was whether different parole decisions were made, inmate by inmate. Would an inmate who would not have been paroled before forecasts were available been paroled after forecasts were available? Conversely, would an inmate who would have been paroled before forecasts were available, not been paroled after forecasts were available? We save the possible impact on recidivism for later in the paper.

3.1 Research Design and Data Collection for the Analysis of Parole Decisions

From July to December 2012, as the operational procedures for providing forecast was being introduced, some cases reviewed by the Board had forecasts available and some did not. Despite best efforts, a subset of parole eligible cases lacked a forecast because requisite data were not ready prior to the date of their parole "interview." Because different inmates had different parole interview dates, and because the forecasting capacity were being gradually assembled, what mattered was the date of an individual's parole interview compared to the date when that individual's forecast could be provided. In addition, a small number of cases had no forecast because their data lacked the required entries for one or more predictors.

Although disappointing from an operational standpoint, whether the forecasts were available provided the opportunity to implement a strong quasi-experimental design. The treatment group had forecasts available when a decision was made about the case. The comparison group did not;

[1]There are good statistical justifications for the how the reliability values were grouped, but a discussion would mean going into considerable detail about the random forests algorithm (Breiman, 2001). For those already familiar with the algorithm, reliabilities can be computed from the "votes" over trees in the random forest. But matters were complicated by there being three prospective outcomes for each individual. It was possible for the winning vote to be a plurality, but not a majority.

forecasts were prepared, but not in time to be used.

Whether the forecasts were available case by case seemed on its face unrelated to features of the case: the background of the inmate, behavior in prison, prison sentence, or prior record. Anecdotally at least, membership in the treatment group or the comparison group seemed much like random assignment. Insofar as random assignment was well approximated, there could be strong internal validity.

How close to random assignment the actual assignment process was can be studied. If the approximation is close, there should be balance in the available variables. That is, the distributions of potential predictors for the treatment group and the comparison group should be very similar. For example, the treatment group and the comparison group should have about the same proportions for the LSIR level, the sex of the offender, and whether there were behavior problems in prison. If sufficient balance can be demonstrated, there can be justification for proceeding as if a real randomized experiment has been undertaken.

3.2 Results

For all of the inmates reviewed during the "burn-in" period, a dataset of 35,842 observations and 51 variables was provided.[2] The first step was to examine the how balanced the treatment and comparison groups really were. There is some technical controversy over exactly how such comparisons should be made (Imani et al., 2008). Statistical tests, for instance, are sample-size dependent and arguably irrelevant.[3] Standardizing the summary statistics can make it difficult to interpret the importance of any apparent differences. It is also unclear how one takes into account the many comparisons made and the correlations between the variables whose balance is being evaluated. For simplicity and ease of interpretation, we consider balance by comparing unstandardized means and proportions.

There is a significant caveat. Even if internal validity is sound, one might well be uneasy about external validity. The data were necessarily collected as the new forecasts were being phased in on an demonstration basis. The intervention being evaluated was not necessarily the intervention that would likely be operational after the demonstration period because the

[2]Many these variable were required to properly organize the data, but were largely irrelevant to the data analysis itself.

[3]With a large sample, the null hypothesis of no difference between the treatment group and the comparison groups can be rejected even if the covariate imbalance makes no material difference in the estimated treatment effect.

immediate goal was to learn how best to introduce and use the forecasts. In short, we are evaluating the impact of the forecasts on decisions made as the intervention was ramping up.

3.2.1 Balance for the Treatment and Comparison Groups

For all inmates, Table 1 compares for the relevant variables available using either the proportion or mean for the treatment group and the comparison group. More complete definitions of the variables can be found in Appendix A. With one exception (in bold font) the summary statistics for the two groups are rather similar, much as you would expect from random assignment. The "violent indicator" variable is problematic. 54% of the treatment grouped were flagged compared to 34% for the comparison group.

The lack of balance for the violent indicator has a credible explanation. Inmates who had been incarcerated for a crime of violence were conventionally flagged because it was thought that such inmates posed a significant threat to public safety. From the forecasting work done for the Board, the researchers knew that the "instant" offense is not a useful predictor when other readily available information is also used, but this had not been fully accepted by the Board at the time the new forecasting procedures were being introduced. Consistent with past practice, inmates with the violent indicator had the highest priority as the new forecasts were being phased in. This is clearly a violation of random assignment to the treatment or comparison groups that can affect the analyses to follow.

Whether the association between the violent indicator and group membership matters depends also on how strongly the violent indicator is related to the Board's parole decisions. In fact, there is only a modest association, which may mean the potential for altering the results is small.[4] Nevertheless, we examine to potential impact of the violent indicator below by reporting separate results depending on whether an inmate is labeled as violent or not. We condition on the violent indicator.

3.2.2 Affects of the Forecasts on the Proportions of Inmates Paroled

From here forward, we focus on inmates called "minimums." "Minimums" are inmates for whom the sentencing judge had set the date for the earliest

[4]When an inmate's conviction crime is violent, parole is granted 52% of the time. When an inmate's conviction crime is not violent, parole is granted 62% of the time. The difference in proportions does not adjust for associations with nonviolent predictors, and is probably overstated.

Table 1: Predictor Balance in Proportions or Means for The Treatment Group and The Comparison Group. ($N = 35,842$)

Predictor	Treatment Group	Comparison Group
Race Black	.46	.46
Race White	.43	.42
Ethnicity Hispanic	.10	.13
Gender Male	.93	.97
High LSIR Level	.55	.56
Medium LSIR Level	.90	.96
Low LSIR Level	.36	.34
Sex Offender	.10	.09
Violent Criminal History in Category 1	.48	.56
Violent Criminal History in Category 1*	.10	.08
Violent Criminal History in Category 2	.17	.16
Violent Criminal History in Category 3	.24	.20
Number of Prison Misconduct Reports	3.4	3.3
Number of Serious Prison Misconduct Reports	1.1	.99
Prison Work Performance Score	2.4	2.5
Prison Behavior Had "Issues"	.09	.10
Number of Prior Arrests	29.5	28.7
Number of Prior Convictions	3.8	5.3
Age at LSIR assessment	35.6	35.3
Age at First Arrest	20.1	20.1
LSIR Score	27.2	27.1
Intelligence Score	90.4	90.5
Guideline Score	4.4	4.7
Nominal Sentence Length in Years	7.2	6.4
Violent Indicator	**.54**	**.34**

mandatory parole review. If such inmates return for a second review, they are no longer called "minimums." Recall that for the three other classes of inmates, the role of the forecasts is very hard to isolate because there has been at least one earlier parole hearing.

62% of the minimums inmates were paroled when the forecasts were not available, and 58% of the minimums inmates were paroled when the forecasts were available ($N = 14,283$). The difference is probably not large enough to matter but with so large sample, the null hypothesis of no difference was rejected. One must also keep in kind that with routine changes in Board membership and natural variation in mix of inmates reviewed, these percentages could change a bit over time or even be reversed.

The following two tables unpack these overall proportions. The proportion of inmates paroled is the outcome of interest. For each table, the top nine rows contain the results when the forecasts were made, but not in time to be introduced into the Board's deliberations. The bottom nine rows contain the results when the forecast were made and available the the Board. The columns headings from left to right are:

1. whether the forecasts were available;
2. the three kinds of forecasts: no arrest, an arrest for a nonviolent crime, or an arrest for a violent crime;
3. the three levels of reliability: low, medium, and high; and
4. the proportion paroled.

When there is an asterisk next to a proportion, a χ^2 test on the table from which the proportions were taken had a p-value of less than .01 (often much smaller) for the null hypothesis of no association.[5] But by and large, the story is to be found in the patterns of proportions.

Table 2 shows the factors related to parole decisions for minimum inmates who were designated as nonviolent because of the nature of the offenses that led to their current sentence. Table 3 shows the factors related

[5]The usual χ^2 test for a contingency table does not take order into account. When there is order in one or more of the variables (e.g., for the forecasts), the test is conservative because it has less power. If the null hypothesis is rejected nevertheless, it would also have been rejected were the ordering built into the test. There is χ^2 test for ordered variables introduced originally by Cochran (1954) and by Armitage (1955) and available in the R library *coin*. We have applied that test when the expected ordering in the results that did not appear. Excellent references are books by Agresti (2002) and Hollander and Wolfe (1999).

Table 2: Minimum Inmates Without the Violent Indicator ($N = 7646$): Proportion Paroled Depending on Whether a Forecast Was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	—	—	Yes	.54*
—	—	Yes	—	—	—	Yes	.51*
—	Yes	—	—	—	—	Yes	.70*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	—	Yes	—	.55*
—	—	Yes	—	—	Yes	—	.59*
—	Yes	—	—	—	Yes	—	.70*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	Yes	—	—	.59
—	—	Yes	—	Yes	—	—	.65
—	Yes	—	—	Yes	—	—	.71
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	—	—	Yes	.39*
Yes	—	Yes	—	—	—	Yes	.52*
Yes	Yes	—	—	—	—	Yes	.73*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	—	Yes	—	.51*
Yes	—	Yes	—	—	Yes	—	.54*
Yes	Yes	—	—	—	Yes	—	.69*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	Yes	—	—	.64
Yes	—	Yes	—	Yes	—	—	.61
Yes	Yes	—	—	Yes	—	—	.73

Table 3: Minimum Inmates With the Violent Indicator ($N = 6637$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	—	—	Yes	.40*
—	—	Yes	—	—	—	Yes	.47*
—	Yes	—	—	—	—	Yes	.58*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	—	Yes	—	.61
—	—	Yes	—	—	Yes	—	.45
—	Yes	—	—	—	Yes	—	.56
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
—	—	—	Yes	Yes	—	—	.60
—	—	Yes	—	Yes	—	—	.55
—	Yes	—	—	Yes	—	—	.54
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	—	—	Yes	.41*
Yes	—	Yes	—	—	—	Yes	.45*
Yes	Yes	—	—	—	—	Yes	.60*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	—	Yes	—	.43*
Yes	—	Yes	—	—	Yes	—	.41*
Yes	Yes	—	—	—	Yes	—	.61*
Forecast Available	Predict None	Predict Nonviolent	Predict Violent	Reliable Low	Reliable Medium	Reliable High	Paroled Proportion
Yes	—	—	Yes	Yes	—	—	.57
Yes	—	Yes	—	Yes	—	—	.58
Yes	Yes	—	—	Yes	—	—	.58

to parole decisions for minimum inmates who were designated as violent because of the nature of the offenses that led to their current sentence. We separate the two because the violent indicators was not in balance; we are conditioning on the violent indicator.

From the last nine rows in Table 2, it appears that the forecasts matter. When the p-value is small, the proportion paroled increases as the forecast changes from an arrest for a violent crime to an arrest for a nonviolent crime, to no arrest. But for this pattern to materialize, the forecasts must have medium or high credibility. These are exactly the sorts of results expected. The same pattern can be seen in the last nine rows in Table 3, which is for the inmates designated as violent, so the presence or absence of the violent indicator does not seem to matter for the pattern of releases.

However, there is a surprise. The similar results can be seen in the first nine rows of Table 2 when the forecasts are *not* available. This is repeated, although not as strongly, in the first nine rows of Table 3. How could that be if the expected pattern is a result of the forecasts?

Recall, that these data were taken from the period when the forecasts were being introduced and before the forecasts became an integral part of the Board's deliberations. Business as usual probably still prevailed, and it was a business as usual in which Board members were required by statute to use their discretion in how different features of each inmate's case were to be weighted. Consequently, the forecasts were seen as additional information provided to the Board, not a replacement for the material that was already used.

Furthermore, the legacy system used by the board was the accumulation of over thirty years of study and work to objectively structure discretionary parole decision making. It already had multiple risk assessment tools that supplemented other information. For many decisions, the forecasts may have been seen duplicative, even if they were seriously considered. For example, whether an inmate had been difficult in prison was included in each inmate's file and was in several forms used in the machine learning forecasts. In summary, despite strong evidence that the machine learning forecasts could in principle improve the Board's decisions, there is no evidence from Tables 2 and 3 that during the burn-in period overall patterns of release changed.

This account can be further examined empirically. Random forests was applied with the decision to parole or not as the outcome. One analysis included predictors from the usual information provided to the Board along with the forecasts and reliabilities. Another analysis included only the predictors from the usual information provided to the Board. (The list of predictors can be found in Appendix A.) Both analyses were able to correctly classify the Board's decisions about 70% of the time and in both cases, the most important variables for classification accuracy were much the same.[6] Behavior while incarcerated and some standard risk indicators dominated. For example, the LSIR was still very consequential.

3.2.3 Changes in Individual Parole Decisions

However, this is not the end of the story. Aggregate patterns may have been much the same, but perhaps, at least at the margin, some inmates who were

[6] With no policy or substantive reason to do otherwise, the costs of false positives and false negatives were given the same weight.

not paroled would have been paroled were the forecasts available, and some inmates who had been paroled would have been paroled were the forecasts available.

There is no way to directly address this question with the data available or any data that could likely be obtained. There is no empirical counterfactual. For example, one cannot see for those inmates reviewed when the forecasts were not available how they would have fared had the forecasts been available. However, one can use the random forest grown when the forecasts were available to *predict* parole decisions for the inmates reviewed when the forecasts were not available for the Board. One simply treats the data for such inmates as a new dataset for which parole decisions are to be predicted.[7]

Table 4 shows the result. 17% of the inmates who had been paroled were projected to not be paroled had the forecasts been available. 48% of the inmates who had not been paroled were projected to be paroled had the forecasts been available. Because these figures are derived from an imperfect statistical approximation of the Board's decision-making, they should not be taken literally. A reasonably circumspect inference is simply that for a substantial number of inmates, a different release decision might well have been made. In addition, there could well be more reversals to grant parole than to deny parole. Were this to be correct, it could help reduce "over-incarceration."

These results may at first seem inconsistent with the earlier conclusion that the forecasts and reliabilities had no demonstrable impact on the proportions of inmates paroled. However, before and during the time the forecasts and reliabilities were introduced, a substantial effort was made to explain the need for the forecasts and how they might be used to make better parole decisions. That educational effort might have encouraged board members to make better use of the usual information provided as well as the new forecasts and reliabilities. Different case-by-case decisions could follow.

It should follow that reversals from parole denials to projected parole releases should have on the average more favorable assessments of risk than reversals from parole releases to projected parole denials. They do. For reversals from parole denials to projected parole releases, 37% were forecasted not to be arrested. For reversals from parole releases to projected parole denials, 30% were forecasted not to be arrested. The difference between

[7]For *all* inmates, forecasts and reliabilities were computed, although some were provided too late for the hearing. This means that the predictors included are the same for inmates reviewed with the forecasts and inmates reviewed without the forecasts.

Table 4: Actual Parole Decision Compared To The Projected Parole Decision For Inmates For Whom the Forecasts Were Not Available (N = 5153)

	Parole Projected	Denial Projected
Actually Paroled	.83	**.17**
Actually Denied	**.48**	.52

30% to 37% comes largely from fewer projected arrests for *noviolent* crime. The projections for violent crime are effectively the same for both kinds of reversals. In other words, reversals from parole releases to projected parole denials, are associated with more forecasted arrests from nonviolent crime. It appears that the introduction of the machine learning forecasts and the usual information largely make the same calls with respect to forecasts of violent crime. Introduction of the machine learning forecasts appears alter some of the Boards decisions for offenders likely to be re-arrested for a nonviolent crime and offenders unlikely to be arrested for any crime.

Even if this interpretation is correct, one cannot conclude that by itself, it was the information provided by the machine learning forecasts that made the difference. The discussions around forecasts may have also gotten members of the parole board to take a much closer look at inmates not having attributes conventionally associated with violent crime. Board members may have soon learned to make useful distinctions between likely nonviolent offenders and offenders who were likely to be crime-free after release. If there is merit in this explanation, it underscores the importance of forecasting different kinds of crime. There is much more to anticipating behavior on parole than whether a parolee is arrested or not arrested regardless of the crime.

4 Impact of the Forecasts on Recidivism

We turn now to the potential impact of the forecasts and their associated reliabilities on re-arrests after release. In order to move beyond the experiences in the burn-in period, we introduce additional data and a regression discontinuity design. At the very least, external validity should be improved.

4.1 Research Design and Data for the Impact on Recidivism

The regression discontinuity design was first proposed by Thistlewaite and Campbell in 1960 (Thistelwaite and Campbell, 1960). Useful elaborations and extensions followed (Trochim, 2001; Imbens and Lemieux, 2008; Berk, 2010) along with some applications in criminal justice settings (Berk and Rauma, 1983; Berk et al., 2010).

In the analyses to follow, inmates who had hearings after the forecasts and reliabilities were available for the majority of inmates become the treatment group. Inmates who had hearings before the forecasts and reliabilities were available for the majority of inmates become the comparison group. The date on which the forecasts and reliabilities became available for the majority can be used as the threshold separating the two groups in time. If the *only* reason why a parole hearing case did or did not have a forecast available was the date of the hearing, one can obtain, in principle, an unbiased estimate (or at least an asymptotically unbiased estimate) of the impact of the forecasts on recidivism. One can have many of the desirable features of an experiment in which the intervention determined is by random assignment. The gradual introduction of the forecasts creates some additional complications to be addressed shortly.

As originally formulated, analyses of regression discontinuity data include all of the observations available, but model specifications are then very consequential. A more recent, alternative formulation compares cases close to and on either side of the discontinuity threshold. This approach, which produces a local estimate of treatment effects, can substantially simplify the statistical analysis, but risks a loss of precision and can limit an findings to the subset of cases near the threshold.[8]

For our RDD analysis, there are four complications. First, it should be clear from our earlier analyses that there actually is no clear threshold. The forecasts were introduced gradually over 6 months, and if there were changes in the Boards deliberations, they materialized gradually over time. One might think that his opens the door for a fuzzy regression discontinuity design (FRDD), but then one must assume that there is a discontinuity in the

[8]Even though the data are longitudinal, and we are estimating a discontinuity on a particular date, the RDD design should not be confused with an interrupted time series design. We have data on individuals so that for any given date, there can be several observations, and the unit of analysis is the individual. For an interrupted time series, design, there is one observation for each point in time, often a summary statistic. For example, a study of the *proportion* of parolee who fail over time might lead to an interrupted time series design. Our data could be organized in that manner, but then a lot of information would be lost.

probability of having the forecasts available (Hahn et al., 2001; Bertanha and Imbens, 2014). Because in our case the intervention was gradually phased in, there is no such break in the probability. We will proceed by trying several reasonable, different threshold dates with the understanding that our treatment effect estimates may be conservative. Some cases assumed to have the forecasts available did not, and some cases assumed to not have the forecasts did.

Second, because we anticipated small treatment effects at best, we could not give up any precision. Therefore, our estimation procedures will use as many observations as possible. In addition, the data are not concentrated near the threshold, but spread rather evenly over the entire range of dates. In short, local estimation methods that focus on observations close to the threshold (Gelman and Imbens, 2014) are effectively off the table.

Third, this decision means that the functional form applied to the date variable becomes very important, and any functional form we might impose must be seen as an approximation of the truth. Nevertheless, asymptotically valid statistical estimates can be obtained for an approximation of the average treatment effect (Berk et al., 2014a; 2014b; Buja et al., 2015).[9] We will emphasis linear functions because of their simplicity and ease of interpretation. But we will explore how robust our findings are by considering several possible nonlinear relationships between date and the response.

Finally, for each observation, the response is categorical. There are three possible outcomes: an arrest for a crime defined as violent crime, an arrest for a crime not defined as violent, and no arrest whatsoever. Had there been only two categories, a generalized regression discontinuity analysis could be applied using logistic regression (Berk and de Leeuw, 1999). With three categories, the generalized RDD applies, but a multinomial logistic regression is required.[10]

For our application, Equation 1 shows the logistic regression formulation for the log of the odds of an arrest for a violent crime ("Violent") compared to no arrest ("None"). Equation 2 shows the logistic regression expression for the log of the odds of an arrest for a non-violent crime ("Nonviolent") compared to no arrest ("None"). In both cases, "Threshold" is an indicator variable coded "0" before the forecasts and reliabilities were available for the majority of inmates, and "1" afterwards. "Dates" is the Julian date for when the parole hearing for each inmate was held. We allow all of the

[9] There are a number of technical issues here that are beyond the scope of this paper.

[10] With categorical outcomes, conventional scatter plots that are otherwise useful in RDD analyses do not provide much visual insight. We do not use them in the analyses to follow.

regression coefficients to differ across the two equations.

The two equations are estimated together to ensure that fitted values behave properly. Because the three probabilities associated with the outcome sum to 1.0, only two equations are required. But a reference category for both must be specified. Our reference category is the absence of an arrest. This choice is made for ease of interpretation. We could have chosen any one of the three categories as the reference and not changed the overall findings.

$$\log\left(\frac{p(y_i = Violent)}{p(y_i = None)}\right) = \beta_0 + \beta_1 Threshold_i + \beta_2 Dates_i. \tag{1}$$

$$\log\left(\frac{p(y_i = Nonviolent}{p(y_i = None)}\right) = \beta_3 + \beta_4 Threshold_i + \beta_5 Dates_i. \tag{2}$$

The data for the pre-intervention cases provisionally includes all parole releases during 2011 and 2012. The data for the post-intervention cases provisionally includes all parole releases during 2013. The relatively few observations before February 2011 were dropped because they were sparsely spread over the month of January.

Both groups had a 24 month follow-up in which any arrests were recorded. Recidivism was defined by the earliest arrest after release and its most serious charge. If there was a charge for a crime defined as violent and a charge for a crime not defined as violent, the former was used to characterize the arrest. Finally, two somewhat different forms of recidivism were included: an arrest while under supervision and an arrest whether under supervision or not.

5 Results

Table 5 shows the multinomial logistic regression results for minimum cases and their arrests while under supervision. There are two comparisons represented: (1) the log-odds of an arrest for a violent crime compared to no arrest, and (2) the log-odds of an arrest for a crime that is not defined as violent compared to no arrest. The values for the coefficients are in logit units. The values for the multipliers are odds ratios, which are produced when regression coefficients for "Threshold" for Equation 1 and Equation 2 are exponentiated. There is no substantive need to include the date variable

Outcome	Coefficient	Multiplier	p-value
Non-Violent Arrest	-0.21	.81	$< .01$
Violent Arrest	-0.42	.66	$< .01$

Table 5: Average Treatment Effect Estimates for Arrests for the "Minimums" While Under Supervision (N = 10,381)

in the table. Its role is apparent in the subsequent figure.[11] The p-values are for the probability of obtaining a regression coefficient for "Threshold" as large or larger than the coefficients obtained if their true values were 0.0. Because the predictors were conceptualized as a random variables, the standard errors were computed with a non-parametric bootstrap (Berk et al., 2014b; Buja et al., 2015). Conventional standards errors provided very much the same results.[12]

Table 5 shows that the odds of an arrest for a non-violent crime are multiplied by a factor of .81 after the forecasts were introduced.[13] The odds of an arrest for a violent crime are multiplied by a factor of .66 after the forecasts were introduced. In both cases, the odds of a re-arrests are reduced. For both effects, one can reject the null hypothesis of 0.0 at well beyond conventional critical levels. Apparently, there are declines in recidivism after the forecasts and reliability became available to the Board.

Figure 1 provides a visual rendering of the results. The vertical axis is in units of odds ranging from 0.0 to .70. The horizontal axis is in units of dates from 2001 to 2014.[14] The role of the date variable is plotted. The outcome of no arrest is the baseline category and is not shown because the information would be redundant.

Over time, there is a modest increase in the odds of an arrest, whether for violent or nonviolent crimes.[15] But when forecasts are made available to the

[11]To be clear, date was included as a regressor. There was just no need to clutter the table with its regression coefficients.

[12]The estimates for which the standard errors are computed are for an approximation of the true regression discontinuity relationship; the approximation is the estimation target. We have asymptotically unbiased estimates of the approximation, not the truth, and the standard errors refer to the approximation as well (Buja et al., 2015).

[13]As mentioned above, some of the very earliest observations were dropped because they were very sparsely distributed.

[14]The dates are transformed back to conventional representations from the Julian dates used in the analysis.

[15]In the units of log-odds, the function of date is linear. In odds units, the relationship becomes non-linear. However, over the empirical ranges of log-odds fitted by the multinomial logistic regression, the amount of non-linearity introduced is very small and difficult

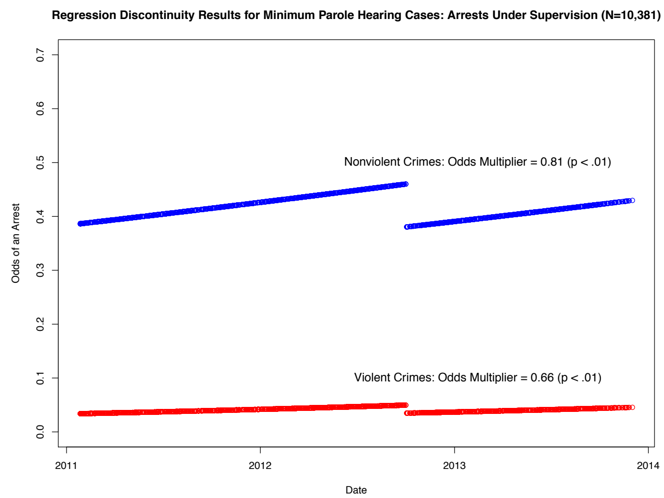

Figure 1: Estimated Discontinuities with No Arrests as the Reference Category (N=10,381) – The Base Percentages were 3.6% Arrested for a Violent Crime And 29% Arrested for Nonviolent Crime

Board, there there appears to be for both kinds of arrests a sharp drop. The decline for violent arrest is smaller to the eye, but that is because the base odds is so much smaller. Immediately before the forecasts are introduced, the odds of a violent arrest compared to no arrest were about .05, or about 20 to 1 against. Immediately before the forecasts are introduced, the odds of a nonviolent arrest compared to no arrest were .48, or about 2 to 1 against. As a result, the multiplier for violent arrests produces a smaller absolute drop in the odds of a violent arrest.

The results from Table 5 and Figure 1 depend on the functional form used for the dates predictor. We added polynomial functions of "Dates" with orders 2, 3, and 4 (i.e., quadratic, cubic and quartic). To avoid some problems raised about this approach (Gelman and Imbens, 2014), the polynomials were entered as B-Splines. This effectively removes the dependence between terms in the polynomial and keeps the units for the higher order terms manageable. The estimated regression coefficients retained their signs and magnitudes. Adding the different polynomials produced AIC measures of fit that did not improve, and the polynomials introduced substantial dependence between the polynomials and the threshold variable. As a result,

to see because of the range of values that must be covered by the vertical axis.

standard error were inflated and p-values soared. There was no evidence that the linear function used was meaningfully in error.[16]

We were also concerned about the timing of the threshold variable because of the gradual introduction of the intervention. Recall that the forecasts and reliabilities began to be provided in July of 2012 and were provided universally by December of 2012. We specified four prospective threshold dates: October 1st, 2012, November 1st, 2012, December 1st, 2012, and January 1st, 2013. For each, a threshold indicator variable was constructed. Each indicator in turn was used in the multinomial logistic regression. All produced results that were easily the same within sampling error.

In is important to emphasize that the estimates of average treatment effects reported are *not* the product of model selection or data snooping. (Berk et al., 2010a; Berk et al., 2014b). No biases were built in because of inductive or adaptive fitting. The linear function was imposed before the data analysis began. Subsequent efforts to try more complex functions of date were undertaken to evaluate the credibility of the linear function imposed.

We repeated the entire analysis using the same three outcomes as before, but for any arrests within 2 years whether under supervision or not. The distribution of arrests did not change much in part because the follow-up interval was relatively short and many parolees who are arrested get into trouble soon after release. It may also be that for many parolees, being under supervision or not had little impact on recidivism.

There were a few more arrests overall. But the main difference was that the number of arrests for nonviolent crimes decreased slightly and the number of arrests for violent crimes increased slightly. Because the most serious crime was used to define whether there was an arrest for a crime was violence, some individuals were re-classified when a later arrest was for a violent crime. Not surprisingly, the overall results and conclusions did not materially change.

6 Summary and Conclusions

There is no evidence that the availability of the forecasts and reliabilities was associated with substantial change in the overall proportion of individuals

[16]We were unable to apply nonparametric smoothers as suggested by Gelman and Imbens (2014). We could find no statistical procedures for smoothers within a multinomial framework. Moreover, had we used smoothers, we risked introducing biases because of model selection when tuning parameters are determined from the data (Berk et al., 2010a).

paroled. After the forecasts were introduced, the overall proportion paroled dropped from 62% to 58%. Given all the factors in play, the 4 percentage point reduction could differ somewhat from year to year or even be reversed. For example, the mix of inmates reviewed can vary over time and membership of the Board can vary over time as well. There was also little change after the forecasts became available in the proportions of inmates released for different levels of forecasted risk and the reliabilities attached. Historical patterns largely remained.

At least part of the explanation is that the standard practices of the Board and the machine learning forecasts were drawing on much the same information. In addition, the forecasts were meant to supplement the information available to the Board, not replace it. Finally, it can be difficult to transform business as usual, at least a first. All members of the Board were experienced and highly credentialed. There was no reason to expect that old ways that had served them well would be rapidly altered.

Yet, there is some evidence from a statistical approximation of the Boards decision-making that availability of the forecasts and reliabilities altered the *mix* of inmates paroled. Arguably, the Board made more accurate distinctions between inmates likely to be arrested for *nonviolent* crimes and inmates unlikely to be arrested for any crimes. This could have altered the decisions made for a substantial number of inmates. Reversals from parole denied to parole granted seem to have been more common that reversals from parole granted to parole denied.

A more careful sorting of "low end" offenders opens a second front in parole policy making. The focus has been on public safety and keeping violent offenders off the streets. But perhaps among inmates who are not likely to be violent, one can do a better job of finding parolees who are very good desistance bets. Just as there can important differences between violent and nonviolent offenders, there can be important differences between nonviolent offenders and offenders very unlikely to re-offend at all.

Recall that had the Board relied exclusively on our forecasts of future dangerousness, recidivism reductions of about 50% were projected. The Board did not rely on the forecasts exclusively for legitimate statutory and administrative reasons. Re-arrests look to have declined by roughly half of the projected reduction after the forecasts and reliabilities were regularly made available.

The reduction in re-arrests was relatively larger for violent crimes than for nonviolent crimes, consistent with the original motivation for the forecasting project. However, the far larger *number* of arrests averted for nonviolent crimes is important too. The consequences for victims are usually less dire, but incarceration of nonviolent offenders is a key driver of mass incarceration practices that can be very costly to offenders and their families.

One reasonably might worry that the substantial reductions in re-arrests are too good to be true because the estimated changes in decisions made by the Board were insufficiently consequential. It is certainly possible that chance factors loaded in a favorable direction. Yet, estimates of impact on Board decisions were made for the period in which the forecasts and reliabilities were just being introduced. In contrast, the data for the analyses of possible recidivism reductions came from a much longer interval both before and after the forecasts became routinely available. They started well before and ended well after the burn-in period. Perhaps that allowed for more credible estimates of changes in recidivism.

Still, one of the weaknesses of the regression discontinuity design is that the average treatment effect estimates can be affected by other interventions introduced around the same time as the treatment. We have no knowledge of such events for this project, and none were ever mentioned in our many discussions with the Board members, parole board staff, and officials from the Department of Corrections. But the practices of the Department of Corrections and the Parole Board are moving targets, and the ways parole supervision are undertaken are often in flux as well. It is possible that some mix practices changed around the time when the forecasts and reliabilities were introduced inflating the treatment effect estimates.

In summary, there is no evidence that the forecasts and reliabilities jeopardized public safety. There is some evidence that they improved it. There is also some evidence that smarter decisions were being made about nonviolent inmates. That may be the most important finding.

References

Agresti, A., (2002). Categorical data analysis. (NewYork: Wiley)

Armitage, P. (1955) “Tests for linear trends in proportions and frequencies.” Biometrtics, 11(3): 375–386

Berk, R.A., (2010) Recent perspectives on the regression discontinuity design. (In A. Piquero and D. Weisburd (eds.) Handbook of quantitative criminology, (New York: Springer)

Berk, R.A. (2012) Criminal justice forecasts of risk: A machine learning approach. (New York: Springer)

Berk, R.A., Barnes, G., Alhman, L., & Kurtz, E. (2010b) When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. Journal of Experimental Criminology, 6(2): 191–208

Berk, R.A., & Bleich, J. (2013) Statistical procedures for forecasting criminal behavior: A comparative assessment. Journal of Criminology and Public Policy, 12(3): 515–544

Berk, R.A., Brown, L., & Zhao, L. (2010a) Statistical inference after model selection. Journal of Quantitative Criminology, 26(2): 217–236

Berk., R.A., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2014a) Valid post-selection inference. Annals of Statistics, 41(2)

Berk, R.A., Pitkin, E., Brown, L., Buja, A., George, E., & Zhao, L. (2014b) Covariance adjustments for the analysis of randomized field experiments. Evaluation Review, 37, 170-196, 2014b)

Berk., R.A., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014c) Misspecified mean function regression: making good use of regression models that are wrong. Sociological Methods and Research, 43: 422–451.

Berk, R.A., & de Leeuw, J. (1999) An evaluation of california’s inmate classification system using a generalized regression discontinuity design. Journal of the American Statistical Association, 94(448): 1045–1052.

Berk, R.A., & Hyatt, J. (2015) Machine learning forecasts of risk to inform sentencing decisions. The Federal Sentencing Reporter, 27(4): 222–228.

Berk, R.A., & Rauma, D. (1983) Capitalizing on nonrandom assignment to treatments: a regression discontinuity evaluation of a crime control program. Journal of the American Statistical Association, 78(381): 21–27.

Bertanha, M & Imbens, G.W. (2014) External validity in fuzzy regression discontinuity designs. National Bureau of Economic Research, working paper 20773.

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

Buja, A., Berk, R.A., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., & Zhang, K. (2015) Models as approximations — a conspiracy of random regressors and model violations against classical inference in regression. $imsart-sts\ ver.2015/07/30 : Buja\_et\_al\_Conspiracy-v2.texdate : July\ 23, 2015$.

Borden, H.G. (1928) Factors predicting parole success. Journal of the American Institute of Criminal Law and Criminology, 19: 328–336.

Burgess, E.M. (1928) Factors determining success or failure on parole. (In A.A. Bruce, A.J. Harno, E. Burgess and E.W. Landesco (eds) The Working of the Indeterminate Sentence Law and the Parole System in Illinois. Springfield: State Board of Parole, pp. 205–249.)

Duwe, G (2014) The development, validity, and reliability of the Minnesota screening tool assessing recidivism risk (MnSTARR). Criminal Justice Policy Review, 25 (5): 579–613.

Cochran, W.G., (1954) Some methods for strengthening the common χ^2 tests. Biometrics, 10(4): 417–451.

Gelman, A., $ Imbens, G. (2014) Why high-order polynomials should not be used in regression discontinuity designs. (No. w20405). Cambridge, MA: National Bureau of Economic Research.

Gottfredson, S.D., & Moriarty, L.J. (2006) Statistical risk assessment: old problems and new applications. Crime & Delinquency, 52 (1): 178–200.

Hamilton, Z., Kigerl, A., Campagna, M., & R. Barnoski, R. (2016) Designed to fit: the development and validation of the STRONG-R recidivism risk assessment. Criminal Justice and Behavior February, 43 (2): 230–263.

Hahn, J., Todd, J.P. & Van der Klaauw, W. (2001) Identification and estimation of treatment effects with a regression discontinuity design. Econometrica, 69: 201–209.

Harcourt, B. (2008) Against prediction: profiling, policing, and punishing in an actuarial age. (Chicago: University of Chicago Press).

Hollander, M., & Wolfe, D.A. (1999) Nonparametric statistical methods, second edition. (New York: Wiley).

Holsinger, A.M. (2013) Implementation of actuarial risk/need assessment and its effect on community supervision revocations. Justice Research and Policy, 15(1): 95–122.

Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal Inference. Journal of the Royal Statistical Society, Series A 171: 481–502.

Imbens, G., & Lemieux, T. (2008) Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142: 611–614.

McCafferty, J.J. (2015) Professional discretion and the predictive validity of a juvenile risk assessment instrument: Exploring the overlooked principle of effective correctional classification. Youth Violent and Juvenile Justice, December 28.

Miller, J., & Malony, C. (2013) Practitioner compliance with risk/needs assessment tools: A theoretical and empirical assessment. Criminal Justice and Behavior, 40(7): 716–736.

Monahan, J., & Skeem, J.L. (2014) Risk redux: The resurgence of risk assessment in criminal sanctioning. Federal Sentencing Reporter, 26(3): 158–661.

Pew (2011) Risk/Needs assessment 101: Science reveals new tools to manage offenders. PEW Center on the States, Public Safety Performance Project. www.pewcenteronthestates.org/publicsafety.

Starr, S.B. (2014). Evidence-Based sentencing and the scientific rationalization of discrimination. Stanford Law Review, 66: 803–872.

Thistlewaite, D.L., & Campbell, D.T. (1960) Regression-Discontinuity analysis: An alternative to the ex-post facto design. Journal of Educational Psychology, 51: 309–317.

Tonrey, M. (2014) Legal and ethical issues in the prediction of recidivism. Federal Sentencing Reporter, 26(3): 167–176.

Trochim, W.M.K. (2001) Regression discontinuity design. (In N.J. Smelser and P.B Bates (Eds.) International Encyclopedia of the Social and Behavioral Sciences, volume 19: 12940–12945).

Appendix A: Information Available to be Board and Used as Predictors

1. Violent Indicator – A binary classification coded "Yes" for a violent offense and "No" otherwise, based on an inmate's offenses that led to his or her current prison sentence.

2. OVRT – A classification into one of four categories called Offender Violent Recidivism Typology that incorporates criminal history into expectations of future recidivism.

3. LSIR Score – A risk assessment score from a Level of Service Inventory-Revised interview that is part of the Pennsylvania Parole Guidelines

4. LSIR Level – Label for risk level given assessment by LSIR instrument

5. Sex Offender – A "Yes" or "No" indicator based upon the Pennsylvania Parole Guideline assessment derived from the Static-99 instrument.

6. Institutional Program Code – A numeric code for prison program participation recorded on the Parole Guideline instrument.

7. Institutional Behavior Code – A numeric code for the offender's behavior and prison adjustment.

8. Guideline Score – A numeric score derived from summing assessment values in the Pennsylvania Parole Guidelines instrument. When the sum exceeds 7, there is a low likelihood of a recommendation for release.

9. Guideline Recommendation – A threshold value of the likelihood of granting parole, as a summation of the Parole Guideline assessment recommendation.

10. Degree of Reliability – One of three possible reliability ranges for the random forests forecasts: greater than 0.5 (a strong result), between 0.5 and 0.4 (a modest result) and less than 0.4 (a weak result).

11. Forecast – The outcome forecasted by random forests: V (violent crime), O (nonviolent crime) or N (no future arrest).

12. Prior Charges – The total count of arrests reported in the rap sheet from Pennsylvania State Police.

13. First Age – The offender's age for the reported first arrest in the offender's criminal history.

14. Arrests – The total number of unique arrest dates in an offender's criminal history record.

15. Sex – A binary code for gender of the offender.

16. LSIR Age – The chronological age of the offender at the time that the LSIR assessment interview conducted immediately before to the parole interview.

17. ISIR Score – The total score from an interview conducted with an inmate prior to the parole hearing.

18. LSIR 29 – The "Yes" or "No" for question 29 in the LSIR assessment pertaining to whether the offender lived in a high crime neighborhood.

19. Convictions – A numeric count for the number of convictions reported on rap sheets manually determined by parole officers.

20. Intelligence Rate – A Department of Corrections intelligence score after a year in prison based upon a group assessment technique.

21. Program Participation – A Department of Corrections rating of institutional program participation after a year in prison.

22. Participation Rating – A Department of Corrections rating of offender work participation after a year in prison.

23. Nominal Length – The computed length of time sentenced based upon the Department of Corrections commitment date and the offender's sentence maximum date.

24. Serious Misconduct – A count to the number of prison misconduct reports withthe most serious misconduct category indicated.

25. Misconduct Counts – A count of the total number of prison misconduct reports found in the offenders complete record.

26. Forecast Printed – whether the forecasts were available to the Board.