UNIVERSITY of PENNSYLVANIA

Department of Criminology

Working Paper No. 2016-6.0

An Evaluation of the Impact of Machine Learning Forecasts on Board Decisions and Recidivism

Richard Berk
Carina Isabel Fink

An Evaluation of The Impact of the Machine Learning Forecasts on Board Decisions and Recidivism*

Richard Berk
Department of Criminology
Department of Statistics
University of Pennsylvania

Carina Isabel Fink
Department of Criminology
University of Pennsylvania

8/23/2016

1 Introduction

The purpose of this report is to evaluate the impact introducing machine learning forecasts of "future dangerousness" into the deliberations of the Pennsylvania Board of Probation and Parole. (Hereafter the Board). The forecasting capabilities were developed using the machine learning procedure random forests (Breiman, 2001), which has been shown to work well in criminal justice applications (Berk, 2012). Unknown was whether the forecasts make a demonstrable difference in practice.

Beginning in 2010, random forests (Hastie et al., 2009, Chapter 15; Berk, 2012) was applied to training data provided by the Board (Board data used for forecast model development) in cooperation with the Pennsylvania Department of Corrections. Full development of the proposed forecasting pro-

*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.

cedures took several iterations as new data were made available and as the Board provided feedback on early results. (A "violent forecast model" was developed using 15 predictor variables identified in the Board training data.) 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 forest forecasts on a routine basis could be provided to the Board in a format that was easy to understand.

The development process for a violent forecast model 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 decision outcome data were collected on the decisions made and how the released individuals fared on parole. An evaluation of the demonstration exercise is the focus of this report.

Three possible parole outcome results from the random forest classification were forecasted : (1) an arrest for a crime defined as violent, (2) an arrest for a crime not defined as violent, and (3) no arrest. Arrests were defined according to Pennsylvania State Police records. Parole Agent violation arrests while under parole supervision and arrests whether under supervision or not were to be considered separately.

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

Three evaluation concerns followed.

1. Did the number of parole releases change because of the forecasts introduced into the Board's deliberations?

2. Did the mix of parole releases change because of the forecasts introduced into the Board's deliberations?

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

All three questions were addressed for different classes of inmates whose circumstances with respect to parole can be rather different because of the

risk that they represent to society and the inmates case history. Since the 1941 Parole Act, the Pennsylvania parole authority determines which inmates can be released at their first eligibility minimum sentence date that was set by a Judge during sentencing. Imprisonment ends at the Court decided maximum sentence date. As a discretionary parole process, decision making for minimum sentence cases has evolved with parole decision-making guidelines during the past quarter century. Formal Parole Guidelines structure release considerations and classify minimum cases according to risk assessment categories and offense types. The decisional platform is the parole interview.

The new forecast score is complementary risk assessment information intended to inform decision-making regarding potential violence and future recidivism. The research goal was to score all minimum sentence eligible cases as one category and separately score all non-minimum cases, which were previously interviewed and denied release, and inmate violators being considered for re-parole. The separation of inmates into policy relevant parole consideration groups is defined by the type of parole interview conducted and a determination of how the new violence forecast score impacts each inmate category during decision making.

The violent offender has been a historic concern of the community and the legislature. In 2008, the Legislature enacted Act 81 to further refine minimum interview types into non-violent sub-categories to insure that a parole is the least restrictive possible for the less violent offender according to their conviction (instant) offense. Two new subcategories were created for non-violent offense minimum interview cases to reduce minimum sentences for qualified minimum cases and/or to expedite parole consideration for the least dangerous offender. The two new minimum interview types were Recidivism Risk Reduction Incentive (RRRI) and Presumptive Parole minimum cases. These interview types engaged both the Courts and the Department of Corrections in the inmate screening process, enabling the Board to focus on the remaining minimum interview types that are more "dangerous" based upon instant offenses that are more serious and violence prone. As a result, the interview type is a mechanism that separates inmates into categories with distinctive levels of presumed risk and different screening procedures. This evaluation examines the impact of the new violence forecast scores on different interview types with respect to their prospects for parole.

Figures **??** and **??** are official statements about the Recidivism Risk Reduction Incentive (RRRI) initiative that clarify the offense makeup of Act 81 inmate interview types for the less serious non violent minimum

sentence interviews.

2 Impact of the forecasts on Board Decisions

The purpose of the random forrest 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 possible 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 that, in turn, provides information about the reliability of the forecasts. The reliability score can range from 0.0 to 1.0, with 0.0 meaning not reliable 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 greater than .5 was considered high, and between .4 and .5 was considered moderate.[1]

The key decision made by the Board is whether to release an inmate on parole. Consequently, one 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. In other words, unless the forecast had sufficiently reliability, it would be effectively ignored. A goal of the evaluation was to determine whether these expectations were borne out.

An additional issue was whether the mix of inmates changes. That is, forecasts of re-arrestes coupled with substantial reliability could also affect the *kinds* of inmates who are paroled, not just their sheer numbers. For example, the importance of the crimes for which an inmate was serving time might be discounted relative to past practice. Such matters are addressed, but a little later.

2.1 Research Design and Data

As the operational procedures for providing forecast was being introduced, some cases considered by the Board had forecasts available and some did not. Despite best efforts, a subset of parole eligible cases lacked a forecast because

[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. Upon request, we are happy to provide that detail.

Figure 1: Summary of the Recidvism Risk Reduction Incentive

Recidivism Risk Reduction Incentive (RRRI) Summary

I. **What is RRRI?** Act 81 of 2008 created a new Chapter 53 in Title 44 that allows a limited class of state prisoners to receive an alternative minimum sentence (known as a "RRRI minimum"). By completing DOC programs designed to reduce their recidivism risk, they are eligible for parole sooner.

II. **Who is eligible for RRRI?** A convicted defendant meeting **all** of the following:
   a. Committed to **custody of DOC;**[1]
   b. No **"history of present or past violent behavior**;"[2]
   c. Not **awaiting trial/sentencing** on charges listed here;[3]
   d. No **weapons** offenses:
      i. Deadly weapon (sentence enhancement, guilty or convicted of offense involving a deadly weapon);
      ii. Offenses listed in Chapter 61 of Title 18; or
      iii. An equivalent offense of other jurisdiction;[4]
   e. Not been convicted/found guilty/adjudicated delinquent of any of the following (or an equivalent offense in another jurisdiction):
      i. **Personal injury crime**[5]—act, attempt, or threat to commit:
         1. **Homicide offenses in Ch. 25** (murder, manslaughter (voluntary/involuntary), causing/aiding suicide, and drug delivery resulting in death);
         2. **Assault and related offenses in Ch. 27** (simple assault, aggravated assault, assault by prisoner, aggravated harassment by prisoner, assault by life prisoner, REAP, terroristic threats, propulsion of missiles into occupied vehicle on a roadway, discharge of firearm into an occupied structure, paintball guns/markers, tear/noxious gas in labor dispute, harassment, stalking, ethnic intimidation, assault of sports official, neglect of care-dependent person, unauthorized administration of intoxicant, threat to use weapon of mass destruction, and terrorism);
         3. **Kidnapping and related offenses in Ch. 29** (kidnapping, false imprisonment, interference with custody of children/committed persons, criminal coercion, disposition of ransom, concealment of whereabouts of child, and luring a child into a motor vehicle or structure);
         4. **Sexual offenses in Ch. 31** (rape, statutory sexual assault, IDSI, sexual assault, institutional sexual assault, aggravated indecent assault, indecent assault, and indecent exposure);
         5. **Arson and related offenses** (18 Pa.C.S. § 3301);
         6. **Robbery** (§3701) and robbery of motor vehicle (§ 3702);

[1] 44 Pa.C.S.A § 5303 definition of "eligible offender" and 44 P.S. §5312.

[2] 44 Pa.C.S.A § 5303(1).

[3] 44 Pa.C.S.A § 5303(5).

[4] 44 Pa.C.S.A § 5303(2).

[5] 44 Pa.C.S.A § 5303(3).

Figure 2: Summary of the Recidvism Risk Reduction Incentive (Continued)

7. 18 Pa.C.S. Ch. 49 Subch. B (**victim/witness intimidation/retaliation**, retaliation against prosecutor or judicial official);
8. 30 Pa.C.S. § 5502.1 (relating to homicide by watercraft while operating under influence);
9. The following **Title 75 offenses**:
   a. DUI w/ bodily injury (former §3731 and Ch 38);
   b. homicide by vehicle (§3732);
   c. homicide by vehicle-DUI (§3735);
   d. aggravated assault by vehicle-DUI (§3735.1);
   e. leaving the scene-accidents involving death/injury (§3742 (relating to accidents involving death or personal injury));

ii. **Specific RRRI excluded offenses:**
   1. 18 Pa.C.S.A. § 4302 (incest);[6]
   2. 18 Pa.C.S.A. § 5901 (open lewdness);[7]
   3. 18 Pa.C.S.A. § 6312 (sexual abuse of children);[8]
   4. 18 Pa.C.S.A. § 6318 (unlawful contact with a minor);[9]
   5. 18 Pa.C.S.A. § 6320 (sexual exploitation of children);[10]
   6. 18 Pa.C.S.A. Ch. 76, Subch. C (internet child pornography);[11]
   7. 42 Pa.C.S.A. § 4302 (drug offenses with firearms);[12]

iii. **Megan's Law offenses** (offenses listed 42 Pa.C.S.A. § 9795.1).[13] This duplicates other provisions but adds the following Title 18 offenses involving a minor:
   1. § 5902(b) (prostitution and related offenses); and
   2. § 5903(a)(3)-(6) (obscene/sexual materials & performances);

iv. **Drug trafficking offenses** (35 P.S. 780-113 (a) (14), (30), & (37)) where defendant sentenced under:
   1. 18 Pa.C.S.A. §7508(a)(1)(iii) (50 lbs marij. etc.);
   2. (2)(iii)(100 grams Schedule I or II);
   3. (3)(iii)(100 grams cocaine);
   4. (4)(iii)(100 grams meth);
   5. (7)(iii)(50 grams heroin); or
   6. (8)(iii)(1000 tablets/300 grams MDA, MDMA, etc.).[14]

[6] 44 Pa.C.S.A § 5303(4)(i).

[7] 44 Pa.C.S.A § 5303(4)(ii).

[8] 44 Pa.C.S.A § 5303(4)(iii).

[9] 44 Pa.C.S.A § 5303(4)(iv).

[10] 44 Pa.C.S.A § 5303(4)(v).

[11] 44 Pa.C.S.A § 5303(4)(vi).

[12] 44 Pa.C.S.A § 5303(4)(vii).

[13] 44 Pa.C.S.A § 5303(4)(viii).

[14] 44 Pa.C.S.A § 5303(6).

requisite data were not ready prior to the date of the parole interview. Because different inmates had different parole interview dates, and because the forecasting capacity was 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 the data lacked the required entries for one or more forecast model predictors. These data limitations continued throughout the study.

Although disappointing from an operational standpoint, whether or not the forecasts were available provided the opportunity to implement a strong quasi-experimental evaluation design. The treatment group had forecasts available. The comparison group did not. Moreover, 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.

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. Likewise the treatment group and the comparison group should have about the same means for the intelligence score, guideline score, number of misconduct charges in prison. If sufficient similarity can be demonstrated, there can be justification for proceeding as if a real randomized experiment has been undertaken. To anticipate, for all but one variable examined, the balance was good. In the analyses to follow, we are able to capitalize on this near balance.

2.2 Results

A dataset of 35,842 observations and 51 variables was provided for monthly parole interviews beginning October 2012 and ending July 2014.[2] The first step in the analysis 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.

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

Table 1: Predictor Balance for The Treatment Group and The Comparison Group. (Proportions or Means Shown for up to 35,842 Observations)

Predictor	Treatment Group	Comparison Group
High LSIR Level	.55	.56
Medium LSIR Level	.90	.96
Low LSIR Level	.36	.34
Sex Offender	.10	.09
Race Black	.46	.46
Race White	.43	.42
Ethnicity Hispanic	.10	.13
Male	.93	.97
Prison Misconduct	.09	.09
Violent Indicator	**.54**	**.34**
Number of Prior Arrests	29.5	28.7
Age at LSIR assessment	35.6	35.3
LSIR Score	27.2	27.1
Intelligence Score	90.4	90.5
Guideline Score	4.4	5.7

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, we considered balance by comparing statistics for unstandardized means and proportions.

2.2.1 How Balanced are the Two Groups?

For the treatment and comparison group, Table **??** compares for the relevant variables available using either the proportion or mean for the treatment group and the control group. With one exception (in bold font) the summary statistics for the two groups are very similar, much as you would expect from random assignment. The "violent indicator" variable is problematic. 54% of the treatment group were flagged compared to 34% for the control group.

The lack of balance for the violent indicator has an operational explanation. Inmates who had been incarcerated for a crime of violence historically were flagged because it was thought that such inmates posed a significant

threat to public safety when released. Consistent with past practice, inmates with the violent indicator had the highest priority as the new forecasts were being phased in. These were the cases provided to Board Members for the initial five months. The intent was to give policy makers a first experience of using a forecast during the decision making process.

By convention, Board Members only interview inmates with a violent instant offense. In March of 2013, the Board expanded its procedures to include all decision makers (Hearing Examiners) conducting interviews. Hearing Examiners only interview inmates with non-violent instant offenses (RRRI and Presumptive Parole). Before all decision makers were included, there is clearly a violation of random assignment to the treatment or control group that could 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.[3] Nevertheless, we examine the potential impact of the violent indicator below by reporting separate results depending on whether an inmate is labeled as violent or not.

2.2.2 Are the Proportions of Inmates Paroled Related to the Forecasts?

Overall, 61% of the inmates are paroled when the forecasts are not available, and 58% of the inmates are paroled when the forecasts are available. The difference is probably not large enough to matter but with so large a sample, the null hypothesis of no difference could not be rejected. One must also keep in mind that with routine changes in Board membership and natural variation in mix of inmates reviewed, these percentages could change and even be reversed.

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

[3]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 other predictors, and is probably overstated.

1. whether the inmate was labeled violent based on the instant offense;
2. whether the forecasts were available;
3. the three kinds of forecasts: no arrest, an arrest for a non-violent crime, or an arrest for a violent crime;
4. the three levels of reliability: low, medium, and high; and
5. 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.[4] But by and large, the story is to be found in the patterns of proportions.

Consider first the "minimum" interview inmates ($N = 14,283$). These are inmates for whom the sentencing judge set the earliest date at which time there would be a mandatory consideration of parole. For this subset of inmates overall, 62% of the minimum inmates were paroled when the forecasts were not available, and 58% were paroled when the forecast were available. With so large a sample, one can reject the null hypothesis of no difference, but the practical difference is probably small.

Table **??** shows the factors related to parole decisions for minimum inmates who are designated as violent because of the nature of the instant offenses (convictions) that led to their current sentence. Table **??** shows the factors related to parole decisions for minimum inmates who are not designated as violent because of the nature of the instant offenses that led to their current sentence. We separate the two anticipating that the forecasts could have different effects on parole decisions. For example, if the Board knows that an inmate has a conviction for violence, that knowledge could be confounded with the role of the forecasts. But the two separate tables, control for that possibility.

From the last nine rows in Table **??**, it appears that the forecasts matter. When the p-value is small, the proportion paroled increases as the forecast changes from no arrest, to an arrest for a nonviolent crime, to an arrest for

[4]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 we expected ordering in the results that did not appear. Excellent references are Agresti (2002) and Hollander and Wolfe (1999).

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

Violent	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	—	Yes	.54*
No	—	—	Yes	—	—	—	Yes	.51*
No	—	Yes	—	—	—	—	Yes	.70*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	Yes	—	.55*
No	—	—	Yes	—	—	Yes	—	.59*
No	—	Yes	—	—	—	Yes	—	.70*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	Yes	—	—	.59
No	—	—	Yes	—	Yes	—	—	.65
No	—	Yes	—	—	Yes	—	—	.71
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	—	Yes	.39*
No	Yes	—	Yes	—	—	—	Yes	.52*
No	Yes	Yes	—	—	—	—	Yes	.73*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	Yes	—	.51*
No	Yes	—	Yes	—	—	Yes	—	.54*
No	Yes	Yes	—	—	—	Yes	—	.69*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	Yes	—	—	.64
No	Yes	—	Yes	—	Yes	—	—	.61
No	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 $p < .01$)

Violent	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	—	Yes	.40*
Yes	—	—	Yes	—	—	—	Yes	.47*
Yes	—	Yes	—	—	—	—	Yes	.58*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	Yes	—	.61
Yes	—	—	Yes	—	—	Yes	—	.45
Yes	—	Yes	—	—	—	Yes	—	.56
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	Yes	—	—	.60
Yes	—	—	Yes	—	Yes	—	—	.55
Yes	—	Yes	—	—	Yes	—	—	.54
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	—	Yes	.41*
Yes	Yes	—	Yes	—	—	—	Yes	.45*
Yes	Yes	Yes	—	—	—	—	Yes	.60*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	Yes	—	.43*
Yes	Yes	—	Yes	—	—	Yes	—	.41*
Yes	Yes	Yes	—	—	—	Yes	—	.61*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	Yes	—	—	.57
Yes	Yes	—	Yes	—	Yes	—	—	.58
Yes	Yes	Yes	—	—	Yes	—	—	.58

a violent crime. But for this pattern to materialize, the forecasts must have medium or high credibility. These are the sorts of results expected.

However, there is also a surprise. The expected pattern of results also can be seen in the first nine rows when the forecasts are *not* available. How could that be if the expected pattern is a result of the forecasts?

The legacy system used by the board is the accumulation of over thirty years of study and work to objectively structure discretionary parole decision making. It has multiple risk assessment tools that supplement and complement one another. An explanation for the similar parole patterns is that the forecasts are simply another way to arrive at largely the same parole decisions in the aggregate. That is, the many decisions about individual inmates lead to the same general patterns in the proportions released. We will return to this question shortly and find that there is much more of interest going. The forecasts appear to be having a significant impact on the *mix* of inmates released.

Tables **??** and **??** are a replay of the Tables **??** and **??**, but for a different group of inmates we label "E,P,F,N" because of their codes for interview type in the data. These interview type codes apply to inmates with nonviolent convictions accordng to the screening applicable in the 2008 legislation:

1. Early RRRI Minimums;
2. Presumptively Rebuttable Minimums;
3. Follow up RRRI Minimums; and
4. New Rebuttable Reviews.

Whether or not the forecasts are available for these inmates ($N = 6616$), about 82% are paroled overall. Within the sorts of random variation one would expect, the details of the story are much like those reported for the minimums. The chances of a parole decline with a re-arrest for a violent crime and non-violent crime, especially if the reliability is at least medium. And as before, whether or not the forecasts are available, does not seem to matter much. For inmates with the violent designation, the expected pattern is found when the forecasts are not available but not when they are. But for the inmates with the violent indicator, the sample size is relatively small.

Tables **??** and **??** have the same structure, but for "minimum" inmates who were considered by the Board previously and were not paroled initially ($N = 8792$). A second review interview is a reconsideration interview. We label the "R" because of their interview type code in the data. 49% are

Table 4: For E,P,F,N Inmates Without the Violent Indicator ($N = 5733$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	—	Yes	.84*
No	—	—	Yes	—	—	—	Yes	.76*
No	—	Yes	—	—	—	—	Yes	.89*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	Yes	—	.75*
No	—	—	Yes	—	—	Yes	—	.78*
No	—	Yes	—	—	—	Yes	—	.86*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	Yes	—	—	.84
No	—	—	Yes	—	Yes	—	—	.87
No	—	Yes	—	—	Yes	—	—	.84
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	—	Yes	.71*
No	Yes	—	Yes	—	—	—	Yes	.86*
No	Yes	Yes	—	—	—	—	Yes	.87*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	Yes	—	.78*
No	Yes	—	Yes	—	—	Yes	—	.81*
No	Yes	Yes	—	—	—	Yes	—	.85*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	Yes	—	—	.82
No	Yes	—	Yes	—	Yes	—	—	.83
No	Yes	Yes	—	—	Yes	—	—	.81

Table 5: : For E,P,F,N Inmates With the Violent Indicator ($N = 883$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	—	Yes	.56*
Yes	—	—	Yes	—	—	—	Yes	.57*
Yes	—	Yes	—	—	—	—	Yes	.78*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	Yes	—	.55*
Yes	—	—	Yes	—	—	Yes	—	.57*
Yes	—	Yes	—	—	—	Yes	—	.78*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	Yes	—	—	.70
Yes	—	—	Yes	—	Yes	—	—	.71
Yes	—	Yes	—	—	Yes	—	—	.75
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	—	Yes	.78
Yes	Yes	—	Yes	—	—	—	Yes	.38
Yes	Yes	Yes	—	—	—	—	Yes	.72
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	Yes	—	.67
Yes	Yes	—	Yes	—	—	Yes	—	.80
Yes	Yes	Yes	—	—	—	Yes	—	.76
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	Yes	—	—	.67
Yes	Yes	—	Yes	—	Yes	—	—	.80
Yes	Yes	Yes	—	—	Yes	—	—	.76

Table 6: For R Inmates Without the Violent Indicator ($N = 3705$) : Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	—	Yes	.46*
No	—	—	Yes	—	—	—	Yes	.62*
No	—	Yes	—	—	—	—	Yes	.69*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	Yes	—	.51
No	—	—	Yes	—	—	Yes	—	.58
No	—	Yes	—	—	—	Yes	—	.56
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	Yes	—	—	.48
No	—	—	Yes	—	Yes	—	—	.59
No	—	Yes	—	—	Yes	—	—	.55
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	—	Yes	.44*
No	Yes	—	Yes	—	—	—	Yes	.66*
No	Yes	Yes	—	—	—	—	Yes	.65*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	Yes	—	.59
No	Yes	—	Yes	—	—	Yes	—	.67
No	Yes	Yes	—	—	—	Yes	—	.67
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	Yes	—	—	.68
No	Yes	—	Yes	—	Yes	—	—	.70
No	Yes	Yes	—	—	Yes	—	—	.52

paroled when the forecasts are not available, and 51% are paroled with the forecasts available. With so large a sample, one can reject the null hypothesis of no difference, but the disparity is very small.

For those without the violent instant offense designation, the same patterns appear, but only for inmates forecasted to be arrested for a violent crime. For those with the violent designation, none of the anticipated patterns appear. It appears that forecasts of future dangerousness, whether from the machine learning output or from other information available to the Board, are being used in some very different fashion. This is likely because these types of review interviews focus on how the inmate performed institutional programming subsequent to the first parole denial.

Tables **??** and **??** include four kinds of "V,S,T,U" inmates ($N = 6136$):

1. Parole supervision violators being considered for re-parole at their first eligibility date;

2. Parole supervision violators previously denied but being reconsidered for re-parole at their next eligibility date.

3. Presumptive parole non-violent criminals who are parole supervision violators under reconsideration for re-parole at their first eligibility date.

4. Presumptive parole non-violent criminals who are parole supervision violators previously denied but being reconsidered for re-parole at their next eligibility date.

Whether or not the forecasts are available, about 52% of the inmates are paroled. The expected patterns appear for only three rows Tables **??**. Whatever information the Board is using to determine a release on parole, it is not substantially related to future dangerousness, at least as defined by the three outcome categories.

In summary, it is relatively common to see the chances of parole decline when the reliability is high and there is a forecast for a non-violent or violent crime. But those results do not hold over all four kinds of inmates and often hold when the forecasts could not be taken into account by the Board. The latter is curious but may have a simple explanation. The Board has access to extensive information about each inmate, much of which is used as input for the machine learning forecasts. For example, whether an inmate has been difficult in prison is included in each inmate's file and in several forms is used to make the machine learning forecasts. In broad brush strokes at least, the Board is taking into account a lot of the same information as the forecasts.

Table 7: For R Inmates With the Violent Indicator ($N = 5087$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	—	Yes	.38
Yes	—	—	Yes	—	—	—	Yes	.63
Yes	—	Yes	—	—	—	—	Yes	.39
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	Yes	—	.37
Yes	—	—	Yes	—	—	Yes	—	.46
Yes	—	Yes	—	—	—	Yes	—	.47
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	Yes	—	—	.39
Yes	—	—	Yes	—	Yes	—	—	.46
Yes	—	Yes	—	—	Yes	—	—	.60
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	—	Yes	.36
Yes	Yes	—	Yes	—	—	—	Yes	.54
Yes	Yes	Yes	—	—	—	—	Yes	.59
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	Yes	—	.49
Yes	Yes	—	Yes	—	—	Yes	—	.58
Yes	Yes	Yes	—	—	—	Yes	—	.48
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	Yes	—	—	.40
Yes	Yes	—	Yes	—	Yes	—	—	.45
Yes	Yes	Yes	—	—	Yes	—	—	.43

Table 8: For V, S, T, U Inmates Without the Violent Indicator ($N = 2833$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	—	Yes	.57
No	—	—	Yes	—	—	—	Yes	.67
No	—	Yes	—	—	—	—	Yes	.66
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	—	Yes	—	.56
No	—	—	Yes	—	—	Yes	—	.58
No	—	Yes	—	—	—	Yes	—	.60
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	—	—	—	Yes	Yes	—	—	.70
No	—	—	Yes	—	Yes	—	—	.64
No	—	Yes	—	—	Yes	—	—	.62
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	—	Yes	.43
No	Yes	—	Yes	—	—	—	Yes	.61
No	Yes	Yes	—	—	—	—	Yes	.66
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	—	Yes	—	.47*
No	Yes	—	Yes	—	—	Yes	—	.66*
No	Yes	Yes	—	—	—	Yes	—	.66*
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
No	Yes	—	—	Yes	Yes	—	—	.42
No	Yes	—	Yes	—	Yes	—	—	.52
No	Yes	Yes	—	—	Yes	—	—	.42

Table 9: For V, S, T, U Inmates With the Violent Indicator ($N = 3303$): Proportion Paroled Depending on Whether the Forecast was Available, The Forecasted Outcome, and the Level of Reliability (* means $p < .01$)

Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	—	Yes	.36
Yes	—	—	Yes	—	—	—	Yes	.50
Yes	—	Yes	—	—	—	—	Yes	.38
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	—	Yes	—	.40
Yes	—	—	Yes	—	—	Yes	—	.43
Yes	—	Yes	—	—	—	Yes	—	.31
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	—	—	—	Yes	Yes	—	—	.39
Yes	—	—	Yes	—	Yes	—	—	.40
Yes	—	Yes	—	—	Yes	—	—	.49
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	—	Yes	.39
Yes	Yes	—	Yes	—	—	—	Yes	.74
Yes	Yes	Yes	—	—	—	—	Yes	.39
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	—	Yes	—	.46
Yes	Yes	—	Yes	—	—	Yes	—	.54
Yes	Yes	Yes	—	—	—	Yes	—	.40
Indicator	Available	None	Other	Violence	Low	Medium	High	Proportion
Yes	Yes	—	—	Yes	Yes	—	—	.40
Yes	Yes	—	Yes	—	Yes	—	—	.45
Yes	Yes	Yes	—	—	Yes	—	—	.43

It may not be surprising, therefore, that the proportions of inmates released have similar patterns whether or not the forecasts are available.

This account can be examined empirically. When the parole decision is regressed on the available predictors both with and without the forecasts and certainties, one finds differences in the relative importance of those predictors. In brief, when the forecasts and certainties are included, they capture a substantial piece of the associations the other predictors otherwise have with the parole decision.[5]

However, tables **??** through **??** only address the proportions paroled. They do not address *the mix* of offenders paroled. Perhaps it is the mix of inmates paroled rather than the proportion paroled that is affected by the forecasts. For example, the proportion of inmates paroled who are forecasted with high reliability to not be re-arrested can be about the same, but the backgrounds of those inmates may differ. It may be that when the forecasts are available, better decisions are made about which inmates are low risk.

2.2.3 Does the Mix of Inmates Paroled Change?

There is no way to directly address this question with the data available or any data that could likely be obtained. But there is an indirect approach with the data we have that may be instructive. The approach is to estimate what the parole decisions would have been had inmates with hearings when the forecasts were not available had hearings when they were. In broad brush strokes, this can be addressed in four steps summarized in Figure **??**.

1. For inmates who were reviewed with the forecasts available, grow a random forest that characterizes the parole decisions made.

2. For inmates who were reviewed without the forecasts available, grow a random forest that characterizes the parole decisions made. There are now two random forests, one when the forecasts were used and one when they were not.

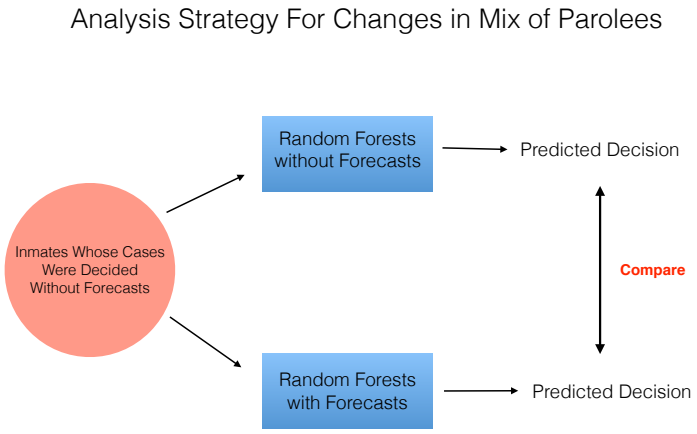
3. Using the inmates for whom the forecasts were *not* available, predict their parole decisions when for forecasts were not available using the random forest from step 2, and predict their parole decisions when the forecasts were available using the random forest from step 1. Note that

[5] There are two logistic regressions each with the parole or refuse decision as the response variable. One regression includes the forecasts and their uncertainties. The other does not. One can then for each of the predictors compare the odds multipliers across the two equations.

for these inmates, the forecast are in the dataset even though they were not available at the time the Board considered the case. Consequently, these inmates can be "dropped" into the the first random forest with no concerns about missing data.

4. Compare the two sets of predictions from the two random forests.[6]

Figure 3: A Statistical Approximation of the Impact of the Forecasts on the Mix of Inmates Released



To build the two statistical procedures, the following predictors were available.

1. Violent Indicator – A sub-classification for the type of offense, such as violent or nonviolent, for the offender's convictions that are used during parole or re-parole consideration.

[6]There are some statistical subtleties involved, but they can be addressed well once one remembers that the treatment group and the comparison group are very similar because of the nearly random way the forecasts were provided. It follows that the two applications of random forests should produce about the same results, except for the impact of the forecasts. Moreover, the fitted class when the forecasts are not available are derived from test data (aka OOB data in random forests). Putting those two facts together, the deck in not being stacked one way or another when predictions for the comparison group are obtained from random forest results based on the treatment group data. In effect, the comparisons are both based on valid test data.

2. OVRT – A classification called Offender Violence 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 PA Parole Guidelines

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

5. Sex Offender – A "yes" or "no indicator based upon PA Parole Guideline assessment using 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 PA Parole Guidelines instrument. If the sum of values exceeds 7, there is a low likelihood of recommended release.

9. Guideline Recommendation – A threshold value of L-likelihood of granting parole, compared to U-unlikely, as a summation of the Parole Guideline assessment recommendation.

10. Degree Of Reliability – A short description for one of three possible statistical forecast score results separated in ranges of over 0.5 (strong result), modest result and low result (lower than 0.4).

11. Forecast – The forecast outcome of the Violence Forecast Model of V (violent crime), O (nonviolent crime) or N (no future arrest).

12. Prior Charges – The Violence Forecast Model parameter indicating the total count of arrests reported in the Rap sheet from PA State Police.

13. First Age – The offenders age for the reported first arrest in the offenders criminal history.

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

15. Sex – A code (M or F) for gender of the offender.

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

17. VFM LSIR Score – The total score of the LSIR interview prior to the parole interview.

18. LSIR 29 – The Y or N score of question 29 in the LSIR assessment pertaining to whether the offender lived in a high crime neighborhood.

19. Convictions – A numeric count of the number of convictions reported on RAP sheets manually ascertained by institutional 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 programming 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 sentence maximum date.

24. Misconduct CAT1 – A count to the number of prison misconduct reports found in the most serious category, "cat1."

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

Consider first the results for the "minimums." For all of the inmates who were reviewed with the forecasts and reliabilities available, all of the variables listed were used as predictors. The outcome was paroled or not. For these inmates, random forests was able to classify 71% correctly. That is, random forests was able to reproduce the true outcome about 71% of the time.[7] The quality of the fit is very respectable, but suggests that some additional factors not included among the predictors are taken into account by the Board. One has a good approximation, but only an approximation,

[7]This is an out-of-sample estimate and not subject to overfitting.

of how the Board makes its decisions when the forecasts are available. One does not have a complete reproduction.

In the second step, the same input were used, except for the forecasts and reliabilities, with the inmates who were considered with the forecasts and reliabilities unavailable. The outcome again was paroled or not. Random forests was able to classify 70% correctly – essentially the same result as before, just as expected.[8] In all other respects, the usual output from the two analyses was much the same.

From the two sets of results, one can predict the parole decisions for the inmates whose hearings did not have the forecasts or reliabilities available, when the forecasts and reliabilities were available and when they were not. The two sets of predictions could then be compared, just as in the Figure. For the minimums, Table **??** shows that 18% of the inmates for whom a refusal was predicted when the forecasts and reliabilities were not available, were predicted to be paroled when they were. 11% of the inmates who were predicted to be paroled when the forecasts and reliabilities were not available, were predicted to be refused when they were. Overall, about 13% of the predicted outcomes differ when the forecasts and reliabilities are available.

Table 10: Minimum Inmates: Changes in the Mix of Parolees (Changed Predictions in Bold Font)

	Forecast Not Available Predict Parole	Forecast Not available Predict Refusal
Forecast Available – Predict Parole	.88	**.18**
Forecast Available – Predict Refusal	**.11**	.82

The analysis was repeated for each of the three inmate groups considered earlier. For the E,P,F,N inmate group, Table **??** shows that 45% of the inmates who were predicted to be refused when the forecasts and reliabilities were not available, were predicted to be paroled when they were. 4% of the inmates who were predicted to be paroled when the forecasts and reliabilities were not available, were predicted to be refused when they were. Overall, about 7% of the predicted outcomes differ when the forecasts and reliabilities are available.

[8]This is also an out-of-sample estimate and not subject to overfitting.

Table 11: E,P,S,N Inmates: Changes in the Mix of Parolees (Changed Predictions in Bold Font)

	Forecast Not Available Predict Parole	Forecast Not available Predict Refusal
Forecast Available: Predict Parole	.96	**.45**
Forecast Available: Predict Refusal	**.04**	.55

For the R inmate group, Table **??** shows that 18% of the inmates who were predicted to be refused when the forecasts and reliabilities were not available, were predicted to be paroled when they were. 13% of the inmates who were predicted to be paroled when the forecasts and reliabilities were not available, were predicted to be refused when they were. Overall, about 16% of the predicted outcomes differ when the forecasts and reliabilities are available.

Table 12: R Inmates: Changes in the Mix of Parolees (Changed Predictions in Bold Font)

	Forecast Not Available Predict Parole	Forecast Not available Predict Refusal
Forecast Available: Predict Parole	.87	**.18**
Forecast Available: Predict Refusal	**.13**	.82

For the V,S,T,U group, Table **??** shows that 32% of the inmates who were predicted to be refused when the forecasts and reliabilities were not available, were predicted to be paroled when they were. 28% of the inmates who were predicted to be paroled when the forecasts and reliabilities were not available, were predicted to be refused when they were. Overall, about 30% of the predicted outcomes differ.

In summary, it appears that the forecasts and their reliabilities complement rather than replace the information already available to the Board. This makes isolating the impact of the forecasts and their reliabilities very challenging. Moreover, there is no direct way to examine what the decisions would have been for inmates reviewed before the forecasts were available had they actually been reviewed when the forecasts were available. We have

Table 13: V,S,T,U Inmates: Changes in the Mix of Parolees (Changed Predictions in Bold Font)

	Forecast Not Available Predict Parole	Forecast Not available Predict Refusal
Forecast Available: Predict Parole	.72	**.32**
Forecast Available: Predict Refusal	**.28**	.68

resorted to a form of statistical simulation.

It appears, nevertheless, that substantial numbers of inmates who in the past would have been paroled are now not paroled. It also appears that substantial numbers of inmates who in the past would not have been paroled are now paroled. Hence, the *mix* of paroled inmates changes significantly even though the forecasts and the reliabilities do not seem to be associated with substantial changes in the *numbers* of inmates paroled.

But, is this change in the mix a good result or a bad result? To address this question we consider the impact of the forecasts and their associated reliabilities on re-arrests after release. To anticipate, the findings are encouraging.

3 Impact of the Forecasts on Recidivism

3.1 Research Design and Data

For estimates of the possible impact of the recidivism forecasts on re-arrests, a regression discontinuity design (RDD) was employed. The regression discontinuity design was first proposed by Thistlewaite and Campbell in 1960 (Thistelwaite and Campbell, 1960). 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., 2010a).

In the analyses to follow, inmates who had hearings after the forecasts were available can be used as a treatment group. Inmates who had hearings before the forecasts were available can be used as a comparison group. As a first approximation, the RDD looks like a simple before-after design. However, before-after designs are weak because it is unclear how comparable the treatment group and comparison group are to begin with. For example,

perhaps the mix of inmates has become more "hard core" over time, and unless that is taken into account, any beneficial effects of the forecasts could be obscured.

With an RDD, one can do better. The *only* reason why a case did or did not have a forecast available was the date of the hearing. Consequently, if one can represent statistically how the hearing date is related to recidivism, any remaining average differences in recidivism between the treatment group and the control group can be attributed to the introduction of the forecasts. In principle, one can have most of the desirable features of an experiment in which the intervention determined is by random assignment.

But there is a price. Compared to a randomized experiment, there can be a substantial loss of precision and hence, statistical power. It more difficult to "find" effects even when they are present. There is also, just as for randomized experiments, the possibility that some other intervention is introduced when the treatment is introduced. What is taken to be an effect of the treatment is actually an effect from something else. We will have more to say about both matters later.

Conventionally, a proper RDD analysis can be accomplished with variants of conventional linear regression. For example:

$$y_i = \beta_0 + \beta_1 Threshold_i + \beta_2 Covariate_i + \varepsilon_i; \tag{1}$$

and

$$\varepsilon_i \sim \text{NIID}(0, \sigma^2), \tag{2}$$

where for the ith case, y_i is the response, $Covariate_i$ is the quantitative variable through which assignment to treatment groups is determined, $Threshold_i$ is a cutoff point on the covariate that separates the treatment group from the control group, and the disturbances ε_i are assumed to be generated independently from a normal distribution with a mean of 0 and a variance of σ^2.

In our setting, cases that fall on or above the threshold value experience the intervention, and cases that fall below the threshold value do not. The threshold is the date on which the forecasts became available. The value of β_1 can provide an estimate of the average treatment effect. Figure **??** is an illustration in which the average value of the response drops by the amount β_1 after the intervention is introduced.[9]

For our RDD analysis, there are three complications. First, there is no guarantee that the relationship between date and recidivism is linear. If the

[9]The figure shows how mean function of regression equation performs. This is what happens on the average.

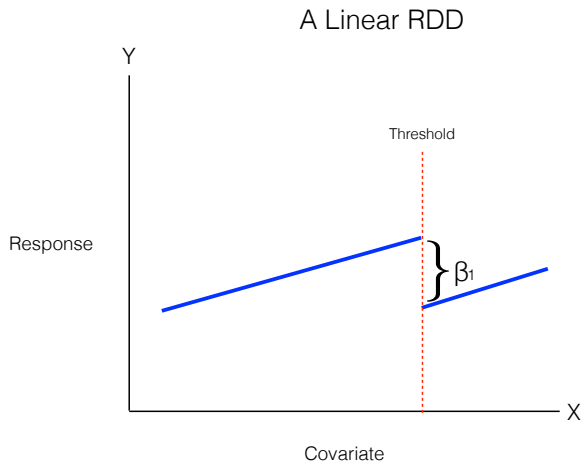


Figure 4: A Linear RDD Analysis with β_1 is the Estimates Average Treatment Effect

relationship is not linear and linearity is imposed, the assumed linearity will likely bias estimate of the average treatment effect. Nevertheless, valid statistical estimates can be obtained with the understanding that the estimates are approximations (Berk et al., 2014a; 2014b; Buja et al., 2015) of the true average treatment effect. Moreover, linearity has its attractions, especially its simplicity and ease of interpretation. We will assume linearity, but will explore how good the approximations are likely to be by considering several possible nonlinear relationships between date and the response.

Second, for each observation, the response is categorical. There are three possible outcomes: an arrest for a violent crime, an arrest for a crime not defined as violent, and no arrest whatsoever. If the outcome had only two categories, the 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 **??** shows the logistic regression equation

[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.

for the log of the odds of an arrest for a violent crime compared to no arrest. Equation **??** shows the logistic regression equation for the log of the odds of an arrest for a non-violent crime compared to no arrest. The right hand side for both has the same linear combination of predictors as for Equation **??**, but we allow for the regression coefficients to differ across the two multinomial logistic regression equations. Both equations are estimated together to ensure that fitted values behave properly.[11]

$$\log\left(\frac{p(y_i = V)}{p(Y_i = N)}\right) = \beta_0 + \beta_1 Threshold_i + \beta_2 Covariate_i. \quad (3)$$

$$\log\left(\frac{p(y_i = O}{p(Y_i = N)}\right) = \beta_3 + \beta_4 Threshold_i + \beta_5 Covariate_i. \quad (4)$$

With three outcome categories, graphs in the spirit of Figure **??** are also more complicated and must conform to Equation **??** and Equation **??**. How this plays out will be explained shortly.[12]

Third, the data are not concentrated near the threshold value but spread rather evenly over the entire range of dates. An important consequence is that local estimation methods that focus on observations close to the threshold (Gelman and Imbens, 2014) are effectively off the table. This puts a heavier burden on how well the relationship between data and recidivism is captured.

The data for the pre-intervention cases includes all parole releases during 2011 and 2012 through October. The data for the post-intervention cases includes all parole releases during 2013. Observations before February 2011 were dropped because they were very 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. That is, 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

[11] For each case, the fitted probabilities for the three outcomes must sum to 1.0.

[12] 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, 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 and a more complicated regression analysis would likely be required. For example, concerns about temporal dependence in the residuals would need to be addressed.

Table 14: Average Treatment Effect Estimates for Arrests for the "Minimums" While Under Supervision (N = 10,381)

Outcome	Coefficient	Multiplier	p-value
Non-Violent Arrest	-0.22	.80	$< .01$
Violent Arrest	-0.42	.66	$< .01$

included: an arrest while under supervision and an arrest whether under supervision or not.

4 Results

Table **??** shows the multinomial logistic regression results for minimum cases and arrests while under supervision. The units of the categorical response are in log odds (aka logits). There are two comparisons represented: (1) the log odds of an arrest for a violent crime compared to the odds of 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 while the values for the multipliers are odds ratios. These are produced when Equation **??** and Equation **??** are exponentiated. There is no substantive need to include the output for the date variable, and in any case, its role is apparent in the subsequent figure.[13]

Table **??** shows that the odds of an arrest for a non-violent crime are multiplied by a factor of .80 after the forecasts are introduced. The odds of an arrest for a violent crime are multiplied by a factor of .66 after the forecasts are 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. Clearly, there are declines in recidivism after the forecasts were introduced.

Figure **??** 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. 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. Overall 3.6% Arrested for a Violent Crime and 29% Arrested for Non-Violent Crime.

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

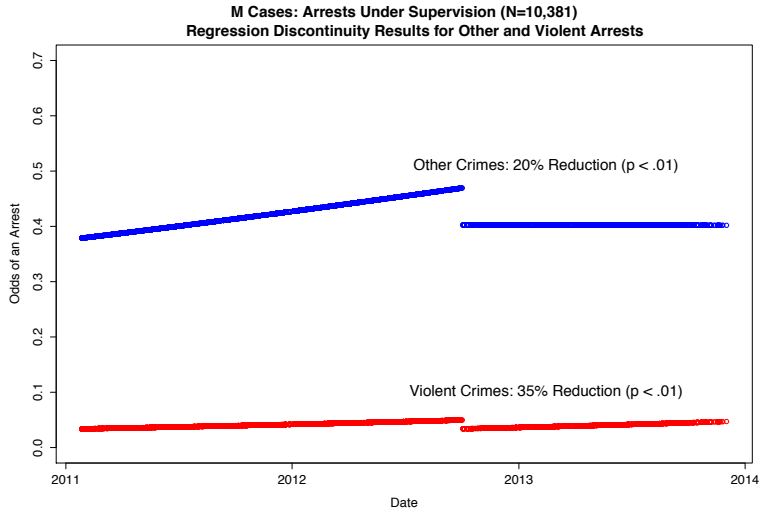


Figure 5: Estimated Discontinuities with No Arrests as the Reference Category (N=10,381) – Overall 3.6% arrested for a Violent Crime, 29% arrested for Non-Violent Crime

Over time, there is a modest increase in the odds of an arrest, whether for violent or other crimes.[14] But when forecasts are made available to the Board, there is a sharp drop in arrests for both kind of crime. The decline for violent arrest is smaller to the eye, but that is because the base odds are so much smaller. Immediately before the forecasts are introduced, the odds of a violent arrest compared to no arrest are about .05 or about 20 to 1 against. Immediately before the forecasts are introduced, the odds of a nonviolent arrest compared to the odds of no arrest are .48 or about 2 to 1. As a result, the multiplier for violent arrests produces a smaller absolute drop in the odds of a violent arrest.[15]

The results from Table **??** and Figure **??** depend on the functional form used for the dates predictor. Recall, that because the data are spread rather evenly over all of the dates, it was not practical to focus just on the cases near the threshold because so many observations would be lost. Having to work with a large range of dates meant that the size of any estimated

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

[15] $.72 \times .05$ is much smaller than $.83 \times .48$.

average treatment would depend heavily on how accurately the relationship between arrests and dates was represented. Consequently, the possibility that the linear function was badly in error was examined in some detail.

1. A smoother was applied using any arrest as the binary response and date as the single regressor. The intent was to consider if the arrest declines started around the time the forecasts were introduced. They did, which is consistent with Table **??**.

2. Using quadratic or cubic functions of date (necessarily in Julian form) altered the estimated average treatment effects a bit but did not improve the AIC fit of the data. Qualitatively, the results were the same, and there was no evidence that a quadratic or cubic function of date was needed.

3. The data were partitioned into two subsets: (1) those cases with either no arrest or an arrest for a violent crimes and (2) those cases with either no arrest or an arrest for a non-violent crime. Equations **??** and **??** were estimated separately using smoothing splines for the function of date within the generalized additive model. Cross-validation was used to determine the complexity of the smoothing spline used for the dates predictor. Again, the results were much the same, and there was no evidence that a function other than linear was needed. The AIC fit was actually worse.

4. The p-values produced by the multinomial logistic software were implausibly small when compared to the p-values associated with the analyses just described. We report the p-values from #3 above which are almost certainly conservative. The correct values are probably much smaller.

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., 2010b; 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.

With the same caveats and diagnostic procedures, the analysis was repeated with recidivism defined by arrests whether the offenders were at the time under supervision or not. The results are much the same as before. Overall 4.9% Arrested for a Violent Crime and 28% Arrested for Non-Violent

Table 15: Average Treatment Effect Estimates for Any Arrests for the "Minimums" (N = 10,381)

Outcome	Coefficient	Multiplier	p-value
Non-Violent Arrest	-0.24	.78	$< .01$
Violent Arrest	-0.37	.69	$< .01$

Crime. Table **??** and Figure **??** show similar declines in re-arrests. In this instance, either set of crime definition produce the very similar declines in recidivism. Diagnostics addressing the linear assumption for the predictor date once again gave no cause for alarm.

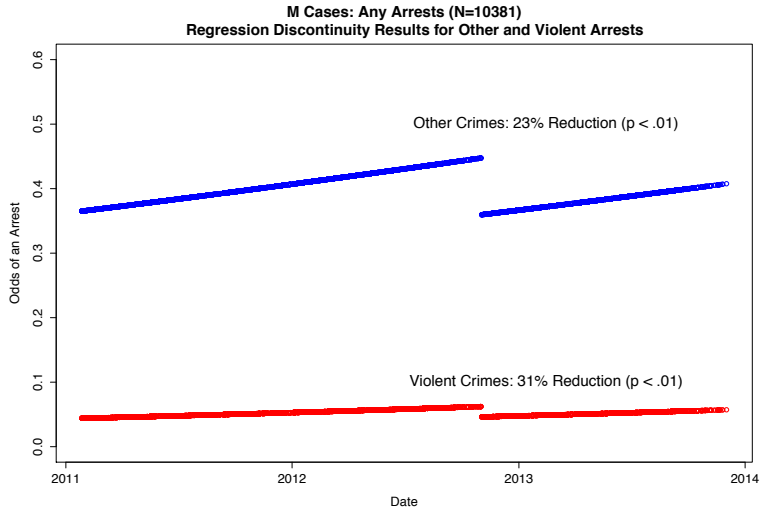


Figure 6: Estimated Discontinuities with No Arrests as the Reference Category (N=10,381) – Overall 4.9% Arrested for a Violent Crime, 28% Arrested for Non-Violent Crime

In addition to the inmates who were getting their sentence-defined, first parole hearing, there was a second group of inmates who were not "Minimums," but also were reviewed by the Board. They were four categories of inmates classified as non-violent offenders as a feature of their sentence. Presumably, they were a smaller threat to public safety than the minimums.

1. Early RRRI minimums — Act 2008-84 (HB7) authorized the Recidivism Risk Reduction Incentive (RRRI) program. Instead of the usual

sentence that might be given, certain non-violent inmates became eligible for an alternative minimum making them eligible for parole sooner.

2. Presumptive Rebuttable Minimums — Act 2008-84 (HB7) also authorized the non-violent offenders meeting the RRRI eligibility criteria would be released on parole at a minimum date unless there were objection from the the court or District Attorney.

3. Follow up RRRI Minimum — Those inmates reconsidered for an RRRI release.

4. New Rebuttable Review — Those inmates reconsidered for a presumptive parole minimum.

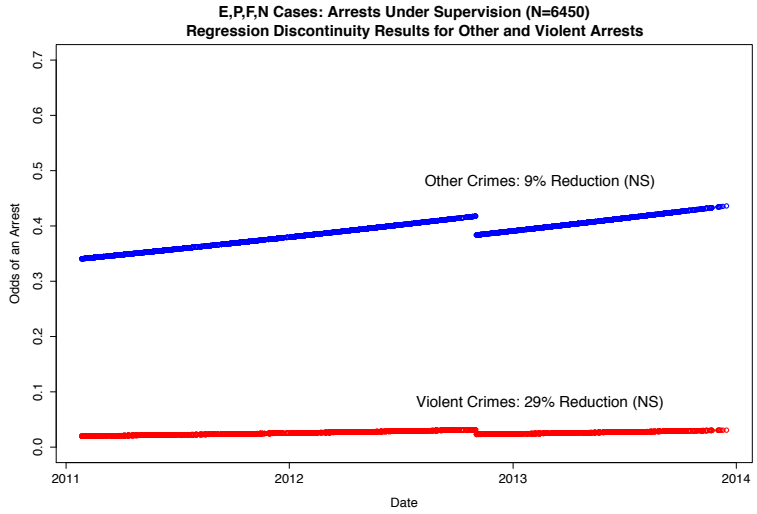


Figure 7: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-violent (N=6450) – Overall 2.5% Arrested for a Violent Crime, 27% Arrested for Non-Violent Crime

In the interest of space, we now just show the results in graphical form. Much as before, Figure **??** and Figure **??** show the results. For the first, 2.5% are re-arrested for a violent crime and 27% are re-arrested for a non-violent crime. For the second, 3.1% are re-arrested for a violent crime and 28% are re-arrested for a non-violent crime. Qualitatively at least, the crime reductions are about same as were found for the minimums, but we cannot at this time reject the null hypothesis that the reductions are actually zero.

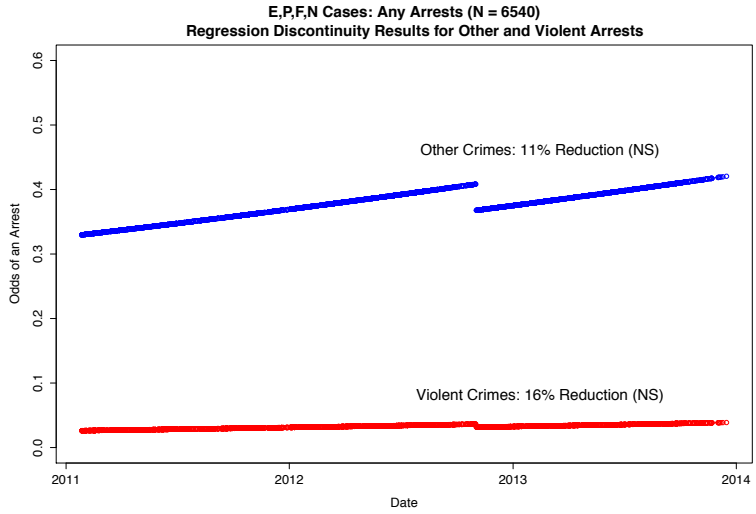


Figure 8: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-Violent (N=6430) – Overall 3.1% Arrested for a Violent Crime, 28% Arrested for Non-Violent Crime

It may be that the substantially smaller number of observations, about 4000 fewer, lack sufficient statistical power. But one can be quite confident that at least the availability of the forecasts did not increase re-arrests for either violent or non-violent crime. However, in this case, different approaches used to get a better handle of how date is related to the outcome bounced the results around substantially.

Figures **??** and **??** are constructed just like the previous four figures, but for the group of inmates who are being considered after having been refused parole earlier. The number of observations is even fewer. For Figure **??**, 5.5% are re-arrested for a violent crime and 35% are re-arrested for a non-violent crime. For Figure **??**, 7.4% are re-arrested for a violent crime and 33% are re-arrested for a non-violent crime. This group of inmates is the most likely to be arrested for violence. Neither change after the forecasts were introduced allows one to reject the null hypothesis of no difference. But the pattern for violent crime is consistent with the earlier five figures. On the other hand, there is again some important variability in the results depending on how the function of date is analyzed.

We use the same format for Figure **??** and **??** with the fewest observations. These figures are for inmates who are being considered for parole having failed on parole previously. For Figure **??**, 4.5% are re-arrested for

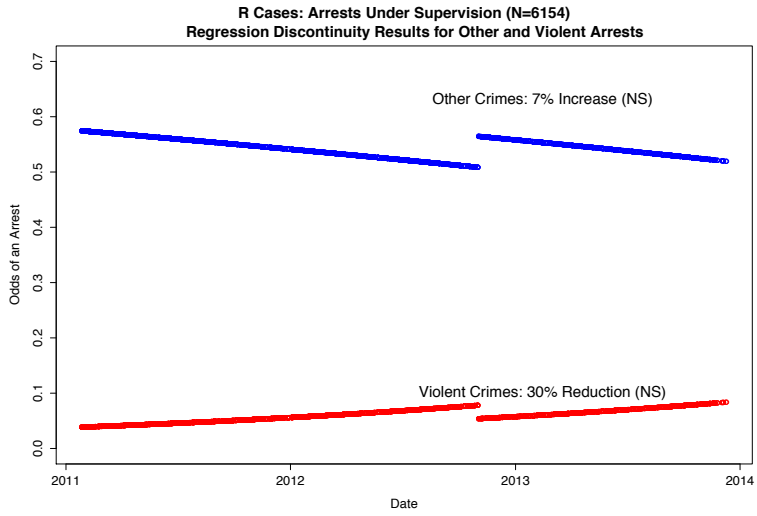


Figure 9: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-violent (N=6154) – Overall 5.5% Arrested for a Violent Crime, 35% Arrested for Non-Violent Crime

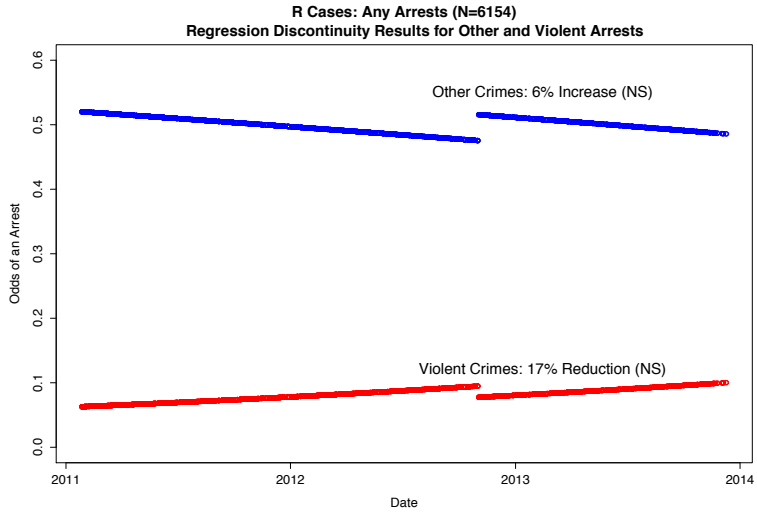


Figure 10: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-Violent (N=6154) – Overall 7.4% Arrested for a Violent Crime, 33% Arrested for Non-Violent Crime

a violent crime and 32% are re-arrested for a non-violent crime. For Figure **??**, 5.8% are re-arrested for a violent crime and 30% are re-arrested for a non-violent crime. Arrests increase after the introduction of the forecasts, but once again, the null hypothesis of no change cannot be rejected.

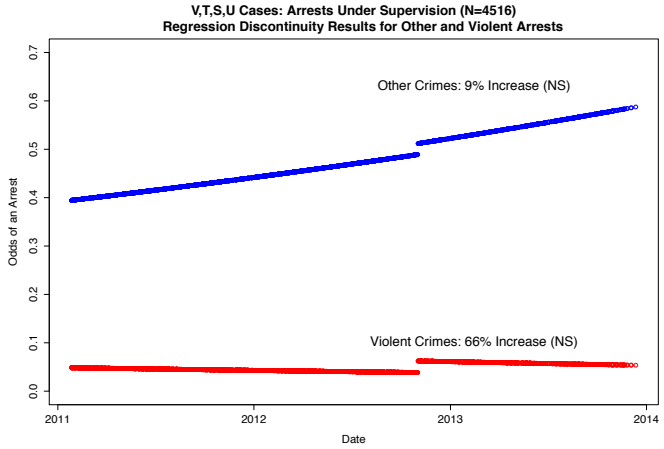


Figure 11: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-violent (N=4516) – Overall 4.5% Arrested for a Violent Crime, 32% Arrested for Non-Violent Crime

4.1 A Somewhat Different Analysis Approach

Even though many of the figures showed meaningful reduction in re-arrests, especially violent crimes, for the majority the null hypothesis of no effect could not be rejected. Indeed, only for the minimum inmates were the results reasonably convincing. And when there were patterns consistent with increases in re-arrests, no compelling conclusions also could be reached.

The data presented significant technical challenges. As mentioned earlier, there is a loss in statistical power inherent in the regression discontinuity design because the threshold variable and the date variable are necessarily correlated. In these data, the variable for threshold was correlated over .80 with the date variable. Key standard errors were inflated by a factor of nearly 4 so that substantial statistical power was sacrificed. Working with squared and cubic functions of date did not improve matters. In the end, it was difficult to reject the null hypothesis even though the sample sizes were relatively large. The implied instabilities meant that relatively small changes in how the relationship between re-arrests and date was represented could

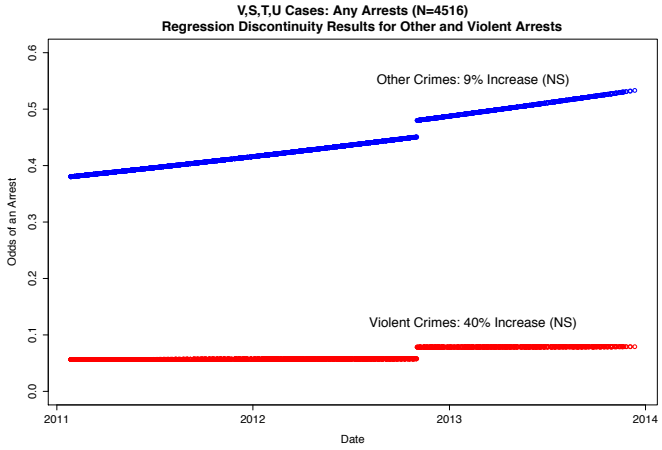


Figure 12: Estimated Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-Violent (N=4516) – Overall 5.8% Arrested for a Violent Crime, 30% Arrested for Non-Violent Crime

alter substantially the size (but usually not the direction) of the estimated average treatment effects.

Coupled with these difficulties is an important policy matter. Although it may be interesting to learn for different classes of inmates how the forecasts may be related to performance on parole, current and future practice would likely have forecasts available for all inmates being considered. It is not clear what the Board would in practice do with different results for different inmate categories.

Putting the technical problems together with the policy complications, we turn to analyses in which all the data are analyzed at once. No distinctions are made between different categories of inmate. We begin with Figure **??**, which shows the smoothed relationship between date and the proportion re-arrested while under supervision for *any* crime. By considering all re-arrests while under supervision, a violent crime is treated the same as a non-violent crime. The good news is that the increasing proportion re-arrested is reversed around the time the forecasts were introduced.[16] The

[16]The word "around" is important. The plot is produced by smoothing so that sharp changes in the proportions over time are intentionally averaged away. The curve looks be decelerating a bit before the threshold, but that could be an artifact of the smoother used (i.e., penalized smoothing splines). Another possibility is the the Board was already beginning to change the way in which various risk factors were treated. There had been a number of discussions with Board members about the key inputs used by random forests.

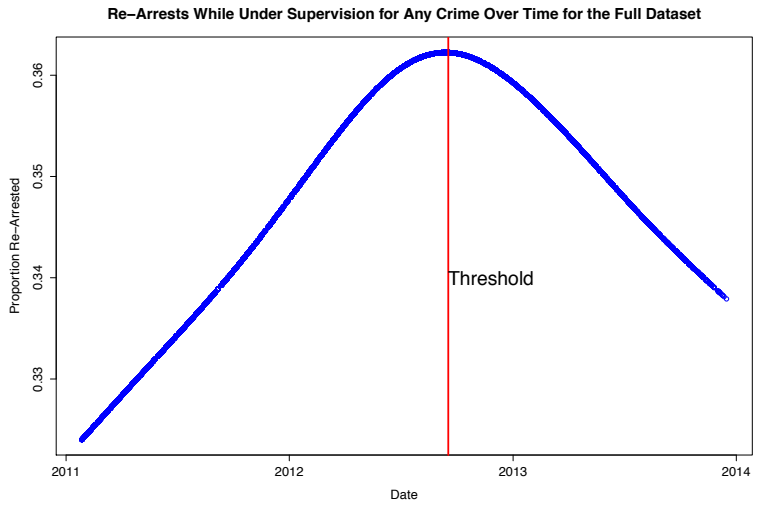


Figure 13: Trends in Re-Arrest While Under Supervision Showing Association With Introduction of the Forecasts (N = 27,646)

bad news is that the changes in the proportions re-arrested are small. The trends are favorable, but hardly exciting.

The results in Figure **??** are necessarily dominated by re-arrests for nonviolent crimes. Violent crimes are a small faction of all arrests. It is important, therefore, to return to the multinomial logistic regression format, but now for all categories of inmate. Figure **??** shows the result. There is a statistically significant reduction in re-arrests for nonviolent crime. The odds of such a arrest are reduced by 9%. Although this may be a modest percentage decrease, the drop in the raw number of re-arrests is more telling. Over the two year follow up period, around 200 arrests were averted.

The decline in re-arrests for violent crime is more dramatic. The odds of re-arrests for violent crime are reduced a statistically significant 34%. Approximation 150 arrests for violent crimes were averted. The percentage reduction is an average result can be larger or smaller depending on the category of inmate.

A range of diagnostic procedures was applied, just as for the earlier analyses. The conclusions were much the same. Statistical dependence between date and threshold was still high, and the results could be altered

For example, the immediate crime of conviction, taken to be a key indicator of risk in the past, was discounted by random forests.

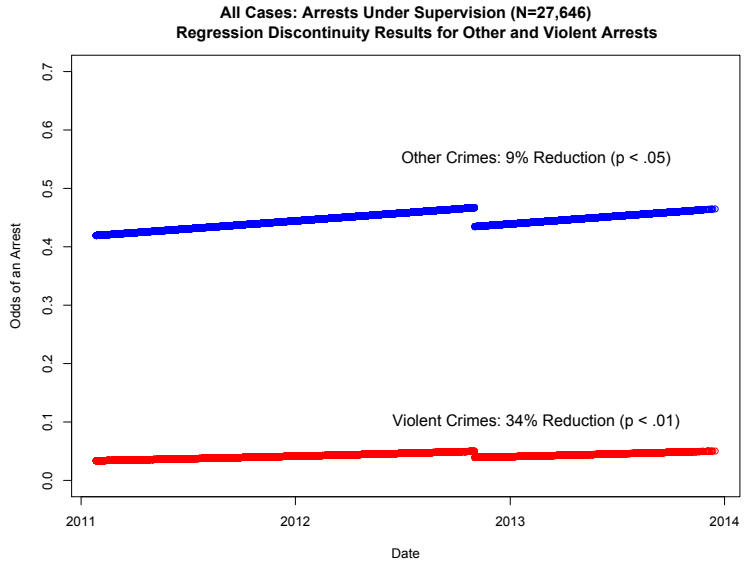


Figure 14: Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-Violent (N = 27,646) – Overall 3.4% Arrested for a Violent Crime, 27% Arrested for Non-Violent Crime

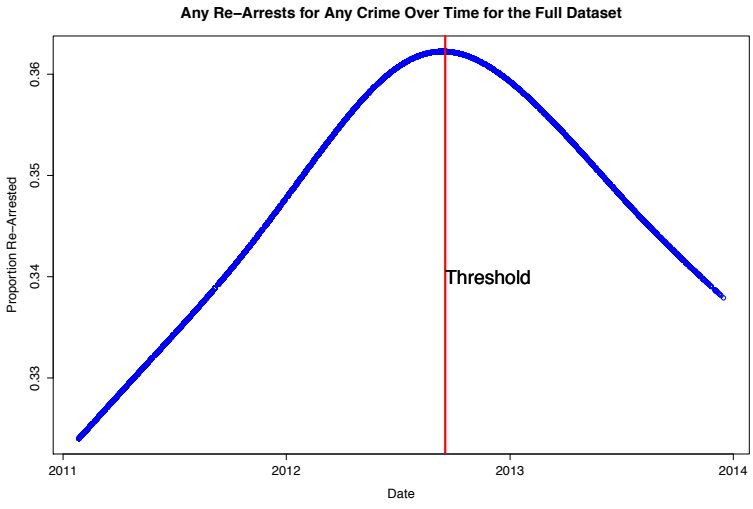


Figure 15: Trends in Re-Arrest While Under Supervision Showing Association With Introduction of the Forecasts (N = 27,646)

by introducing polynomials of date. However, the polynomials and other approaches used to check the mean function never improved the fit (i.e., the AIC) and typically made the fit worse.[17] There was, therefore, no need to consider revising the linear function of date in the multinomial logistic regression.

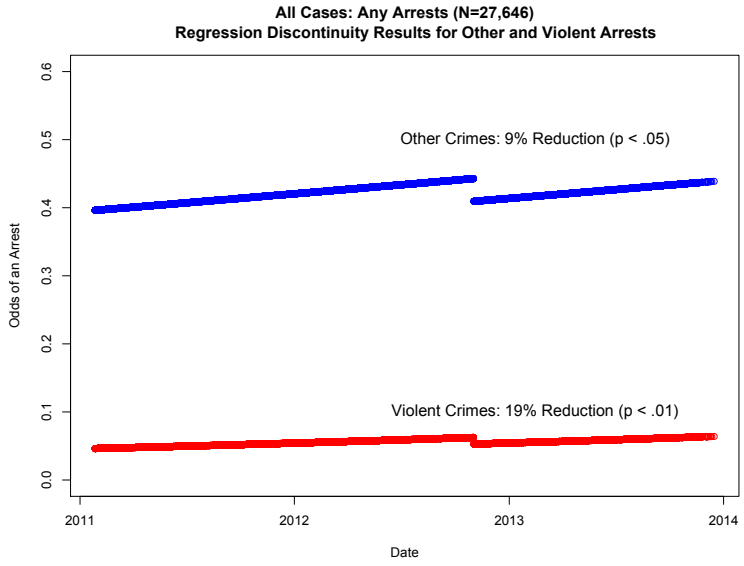


Figure 16: Discontinuities with No Arrests as the Reference Category for Inmates Classified as Non-Violent (N = 27,646) – Overall 4.0% Arrested for a Violent Crime, 31% Arrested for Non-Violent Crime

Figures **??** and **??** repeat the analyses for all re-arrests, not just those while under supervision. The results are qualitatively the same, but the reduction in the odds of a re-arrest for a violent crime is now 19%. Without more data, one can only speculate about the explanation, but it appears that the forecasts are less strongly associated with decreases in re-arrest when the arrests include crime committed while not under supervision. These analyses too passed muster with the same variety of diagnostic procedures.

[17]The residual deviance hardly changed at all even though more degrees of freedom were being used up.

5 Conclusions

Despite all of the details, the overall conclusions are simple. First, there is little systematic evidence that the availability of the forecasts was associated with substantial changes in the number of individuals paroled. After the forecasts were introduced, the overall proportion paroled dropped from 61% to 58%. In percentage terms at least, the drawdown from state prisons and the existing number of individuals on parole were not materially changed.

Second, at least part of the explanation is that the historical practices of the Board and the machine learning forecasts were drawing on much the same information. When the forecasts become available, weight given to predictors conventionally used simply declined. It may also be that at least implicitly there are habitual precedents for the fraction of inmates paroled that help shape the sequence of parole decisions.

Third, there is substantial evidence that the availability of the forecasts altered the mix of inmates paroled across all four categories of inmate. On average, reversals from a predicted refusal to a predicted parole were more likely than reversals from a predicted parole to a predicted refusal. It is important to stress that these conclusions depend on statistical representations of factors shaping release decisions with and without the forecasts. It was impossible to have the same set of inmates actually reviewed by the Board with and without the forecasts being available.

Finally, over the full set of inmate categories, re-arrests declined after the forecasts were made available. The association with the reduction in re-arrests was stronger for violent crime and non-violent crime. This differential impact is fully consistent with the origins of the forecasting project. The primary goal, at least initially, was to reduce the number of parolees who committed violent crimes. A forecast with high reliability that an inmate, if paroled, will be arrested for a violent crime looks to have been very influential.

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." *Handbook of Quantitative Criminology*, A. Piquero and D. Weisburd (eds.), New York: Springer.

Berk, R.A., Barnes, G., Alhman, L., and E. Kurtz (2010a) "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, 2010.

Berk, R.A., (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. New York: Springer.

Berk, R.A., Brown, L., and L. Zhao (2010b) "Statistical Inference After Model Selection. *Journal of Quantitative Criminology*, 26(2): 217–236.

Berk., R.A., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2014b) "Valid Post-Selection Inference." *Annals of Statistics* 41(2).

Berk., R.A., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., and L. Zhao (2014a) "Misspecified Mean Function Regression: Making Good Use of Regression Models that are Wrong." *Sociological Methods and Research* 43: 422-451.

Berk, R.A., and Rauma (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, 1983.

Berk, R.A., and 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., Pitkin, E., Brown, L., Buja, A., George, E., and L. Zhao (2014b) "Covariance Adjustments for the Analysis of Randomized Field Experiments," *Evaluation Review* 37, 170-196, 2014).

Breiman, L. (2001a) "Random Forests." *Machine Learning* 45: 5–32.

Buja, A., Berk, R.A., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and K. Zhang. (2015) "Models as Approximations — A Conspiracy of Random Regressors and Model Violations Against Classical Inference in Regression." $imsart-sts\ ver.2015/07/30 : Buja\_et\_al\_Conspiracy-v2.texdate : July\ 23, 2015.$

Cochran, W.G., (1954) "Some Methods for Strengthening the Common χ^2 Tests." *Biometrics* 10(4): 417–451.

Gelman, A., and G. Imbens (2014) "Why High Order Polynomials Should Not Be Used in Regression Discontinuity Designs." Working Paper, Department of Statistics, Columbia University.

Hastie, T.J., and R. Tibshirani (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Hastie, T., Tibshirani, R. and J. Friedman (2009) *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Hollander, M., and D.A. Wolfe (1999) *Nonparametric Statistical Methods*, second edition. New York: Wiley.

Imai, K., King, G., and E.A. Stuart (2008). "Misunderstandings Between Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society*, Series A 171: 481–502.

Imbens, G., and T. Lemieux (2008) "Regression Discontinuity Designs:A Guide to Practice."' *Journal of Econometrics* 142: 611–614.

Thistlewaite, D.L., and D.T. Campbell (1960) "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Design." *Journal of Educational Psychology* 51: 309–317.

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.