Evaluating Pre-trial Programs Using Interpretable Machine Learning Matching Algorithms for Causal Inference

Travis Seale-Carlisle[*1], Saksham Jain[*2†], Courtney Lee[3], Caroline Levenson[3], Swathi Ramprasad[3], Brandon Garrett[3], Sudeepa Roy[3], Cynthia Rudin[3], Alexander Volfovsky[3]

[1]University of Aberdeen
[2]University of Washington
[3]Duke University
travis.seale-carlisle@abdn.ac.uk, sj305@uw.edu, courtney.lee@duke.edu, cglevenson18@gmail.com, ramprasadswathi@gmail.com, bgarrett@law.duke.edu, sudeepa@cs.duke.edu, cynthia@cs.duke.edu, alexander.volfovsky@duke.edu

Abstract

After a person is arrested and charged with a crime, they may be released on bail and required to participate in a community supervision program while awaiting trial. These 'pre-trial programs' are common throughout the United States, but very little research has demonstrated their effectiveness. Researchers have emphasized the need for more rigorous program evaluation methods, which we introduce in this article. We describe a program evaluation pipeline that uses recent interpretable machine learning techniques for observational causal inference, and demonstrate these techniques in a study of a pre-trial program in Durham, North Carolina. Our findings show no evidence that the program either significantly increased or decreased the probability of new criminal charges. If these findings replicate, the criminal-legal system needs to either improve pre-trial programs or consider alternatives to them. The simplest option is to release low-risk individuals back into the community without subjecting them to any restrictions or conditions. Another option is to assign individuals to pre-trial programs that incentivize pro-social behavior. We believe that the techniques introduced here can provide researchers the rigorous tools they need to evaluate these programs.

Introduction

In the United States, about five-hundred thousand people are held each day in jail while awaiting a criminal trial, often because they cannot afford to pay cash bail (Minton and Zeng 2021). This problem is compounded by a large body of research that suggests that pre-trial detention causes a variety of negative outcomes (Dobbie, Goldin, and Yang 2018; Heaton, Mayson, and Stevenson 2017; Leslie and Pope 2017; Lowenkamp, VanNostrand, and Holsinger 2013; Phillips 2012; Stevenson 2018). In an attempt to improve pre-trial justice, many people including researchers, policymakers, and legal scholars have expressed a strong interest in alternatives to pre-trial detention (Garrett 2022).

A common alternative to pre-trial detention in the United States is community-based supervision (Mahoney et al. 2001). The primary goal of these 'pre-trial programs' is to provide community support for people while they await trial. The programs also aim to reduce re-arrests and failures to appear in court. They typically provide access to education, employment training, and transitional housing as long as certain conditions and restrictions are met (Clarke 1988; Mamalian et al. 2011; VanNostrand and Keebler 2009; VanNostrand, Rose, and Weibrecht 2016). There is often considerable discretion in selecting which community supervision conditions are imposed, but a pre-trial services agency may provide the judge with recommendations concerning what program may be suitable for a person, and that agency may then be responsible for supervising that program. For example, people who take part in these pre-trial programs may need to have periodic check-ins with a case manager, maintain employment, undergo alcohol testing and treatment, wear an electronic monitor, or take part in cognitive behavioral therapy (Levin 2007; Clark and Henry 2003). If they fail to meet the imposed conditions or violate any restrictions, they may have their bond conditions modified, and could be jailed for the duration of the pre-trial period.

While pre-trial programs are common throughout the United States (Mahoney et al. 2001), there is little research demonstrating that such programs are effective at reducing new arrests (Cadigan and Lowenkamp 2011; Cohen 2012; Lowenkamp and VanNostrand 2013; Mamalian et al. 2011; Robinson et al. 2011). Based on limited data available, researchers have found no impact on re-arrest rates (Austin, Krisberg, and Listky 1985; Cooprider and Kerby 1990) or a slight reduction in re-arrest (Goldkamp and White 2006; Lowenkamp and VanNostrand 2013). Bechtel et al. (2017) conducted a meta-analysis of the pre-trial literature to clarify these mixed findings. They emphasized that while the research does not support any finding that pre-trial supervision is effective in reducing re-arrest, the quality of the extant research "was not very good" (p. 460) and there is a "great need for new and more rigorous pre-trial research in all related areas" (p. 459). We answer that call here by introducing a rigorous method for evaluating pre-trial programs.

[*]These authors contributed equally.
[†]Work done while at Duke University

Evaluating Pre-trial Programs

Many consider the randomized controlled trial (RCT) to be a "gold standard" method for program evaluation, because the RCT has the potential to establish a causal link between a treatment and an outcome. However, RCTs are often not viable because of legal, ethical, and logistic constraints. Judicial officers would not, for obvious constitutional and ethical reasons, randomly jail people rather than release them to community supervision. Pre-trial researchers therefore need a program evaluation method that can establish causality in lieu of random assignment. Here, we use approaches for observational causal inference to do that.

Interpretable machine learning techniques for causal inference have only recently become available (Wang et al. 2021; Dieng et al. 2019; Parikh, Rudin, and Volfovsky 2022; Lanners et al. 2023), filling a gap in the observational causal inference literature. Previous machine-learning-based causal inference methods for estimating treatment effects from observational data all suffer from one of three typical flaws. The first flaw occurs when causal inference methods produce estimates of causal effects that are uninterpretable and not able to be manually examined and verified. This includes propensity score matching methods (Rosenbaum and Rubin 1983), which lump data together that are far apart in covariate space, genetic matching (Diamond and Sekhon 2013), as well as black box machine learning methods, e.g. causal forest, prognostic score matching, and targeted maximum likelihood estimation (Chipman et al. 2010; Hill, Reiter, and Zanutto 2004; Wager and Athey 2018; Van Der Laan and Rubin 2006). Without interpretability, where the problem is unsupervised and the ground truth treatment effect is not known, the trustworthiness of results is difficult to assess. The second flaw occurs when causal inference methods require the analyst to manually determine what to match on or determine the bins for matching, as in coarsened exact matching (Iacus, King, and Porro 2012). Such methods are frequently inconsistent between analysts and use humans to perform manual high-dimensional distance metric optimization, a skill at which they are not naturally adept. The third flaw occurs when causal inference methods produce causal estimates that are not very accurate. Linear regression is an example because linear models are not particularly flexible.

In contrast to methods mentioned above, the new method we consider, FLAME-DAME (Wang et al. 2021; Dieng et al. 2019) does not suffer from any of these flaws. It (1) provides interpretable matched groups, which can be scrutinized by domain experts, (2) uses machine learning to automatically identify important variables for matching, and (3) yields accurate causal estimates. Through the use of this new technique, we can examine data from pre-trial programs through a more fine-grained lens than ever before.

Our Work

This work evaluates a typical pre-trial program, administered in Durham, North Carolina, operated by the Criminal Justice Resource Center (CJRC) between 2016 and 2019. Durham is a medium sized, typical jurisdiction in the United States and the CJRC is a government entity located in Durham that provides support and treatment to justice-involved individuals both in jail and in the community (Parmer and Merritt 2019). The CJRC provides supervision in lieu of pre-trial incarceration largely for low-risk individuals. The CJRC supervises people by requiring periodic check-ins with case managers, or, in some cases, requiring people to wear electronic monitors. Those supervised by the CJRC often have access to behavioral health treatment, employment training and placing assistance, housing services, and educational services. Their vision is to support clients' successful reentry into the community, leading to reduced arrest rates and fewer negative consequences that impact communities when individuals cycle in and out of jail (Parmer and Merritt 2019).

We tested whether the CJRC program reduced the probability of new criminal charges. That is, we tested whether those who took part in the CJRC program were less likely to be arrested and charged with a new crime than their similar counterparts who spent the duration of their pre-trial period in the community without any conditions or restrictions on their release.

We used multiple methods of observational causal inference to conduct this test, including propensity score matching (Rosenbaum and Rubin 1983) and the new FLAME-DAME matching algorithms (Wang et al. 2021; Dieng et al. 2019). To preface our results, all methods provided an estimated average treatment effect of approximately zero. However, importantly, the results using FLAME-DAME confirmed the expert recommendations used for variable selection and did not change the conclusions. Thus, the results of this study indicate that there is no way, from this dataset, to conclude that the program is effective at reducing new criminal charges in this community. If this finding replicates across other datasets, policymakers might consider alternative strategies in managing low-risk populations or might consider focusing their resources on higher-risk populations instead. Before making substantial changes to policy, however, we recommend researchers adopt these new techniques to rigorously evaluate pre-trial programs.

Our key contributions include (1) close collaboration with criminal justice experts in understanding, determining, and experimentally validating important covariates that influence pre-trial program selection, and (2) through comprehensive analyses, highlighting the significance of a quasi-experimental approach utilizing interpretable machine learning techniques, for rigorous program evaluation in criminal justice settings.

Data Collection and Preparation

Program Eligibility. The justice-involved individuals in our study were arrested, booked in the Durham County jail, and assessed shortly after by a social staff worker employed by the CJRC. Based on that assessment, the CJRC either recommended individuals to be released and to take part in the pre-trial program or to remain detained in jail. Broadly speaking, low-risk people charged with either a misdemeanor or low-level, non-violent felony were recommended for the pre-trial program. High-risk people charged with a violent felony were rarely recommended for the pre-trial program. A judge

assigned to the case reviewed these recommendations and had considerable discretion to either accept or reject them.

Study Design. This study adopted a quasi-experimental design that compares individuals who were recommended and took part in the program to those who were recommended, released from jail, but did not take part in the program. Participation in the pre-trial program is therefore the *treatment* condition in our study, whereas pre-trial release without supervision is the *control* condition in our study.

Data Collection. We used three separate datasets to compare these two groups of individuals. The CJRC data includes information on the risk score, criminal history, education level, age, gender, ethnicity, and indigency status of individuals. We merged the CJRC data with the data from the Durham County jail to determine bond conditions and confirm whether individuals were released or detained pre-trial. We then merged these data with the data from the North Carolina Administrative Office of the Courts (AOC) Automated Infraction Corrections System (ACIS). The ACIS extract we used largely contained criminal records for cases occurring between January 1st, 2016, to July 31st, 2020. We used the ACIS extract to search for case outcomes, sentencing information, and new criminal charges.

Data Cleaning and Processing. The CJRC recommended 1,787 people for their pre-trial program during 2016-2019. We took several steps to ensure that these data were clean and reliable. First, we used a comprehensive search process to cross-reference the individuals across the CJRC, AOC, and Durham Jail datasets. We searched for matching case record numbers, names, date of births, and other identifying information (e.g., ethnicity and age). If we could not successfully cross-reference individuals, we excluded them from the final merged dataset. After that, we excluded individuals that were convicted and incarcerated in prison because they would be unable to commit additional crimes. Third, there were several individuals that had taken part in the CJRC program on multiple occasions. For these individuals, we only kept their first involvement with the CJRC program in the final dataset. Fourth, we excluded individuals from the final dataset if their case was still pending or was disposed of after July 31st, 2019. This ensured that we had criminal records for everyone in the final dataset for at least one-year after case disposition, which is important for our outcome measure discussed below. Lastly, we excluded individuals if we could not confirm their pre-trial release. After excluding these individuals (n = 372), we retained 1,415 people in our final dataset, of which 687 people took part in the program and 728 people did not. Of those who took part in the program, 562 people completed it. Figure 1 shows a flow chart that depicts this data cleaning process. The data collection and cleaning process is described in further detail in the supplementary material.

Outcome Measure. The outcome in our study is new criminal charges, which we define as any new criminal charge occurring within the one year after case disposition. We chose to measure criminal charges after case disposition instead of during pre-trial in part because of substantial variation in the time it takes to complete the pre-trial program. Some individuals took part in a two-week program and waited several months for their case to be disposed, whereas others took part in a three-month program which ended several days prior to their case disposition. Our measure of new criminal charge avoids this issue. Second, we chose a one-year window because we had statewide criminal records for every person for the one year after their case disposition date. If we chose a larger window than this, we would need to remove individuals from the dataset. Lastly, no one in our final dataset was incarcerated in prison after their case was disposed. However, there were 166 people that were either incarcerated in jail or required to do community service for a brief portion of the one-year after their case disposition. The ACIS extract we had access to did not distinguish between those two sentencing outcomes and instead grouped them together. Removing those 166 people from the dataset and repeating our analyses (discussed in subsequent sections) does not change any of our results or conclusions. The remaining 1,249 people in our final dataset had their charges dismissed or were not convicted.

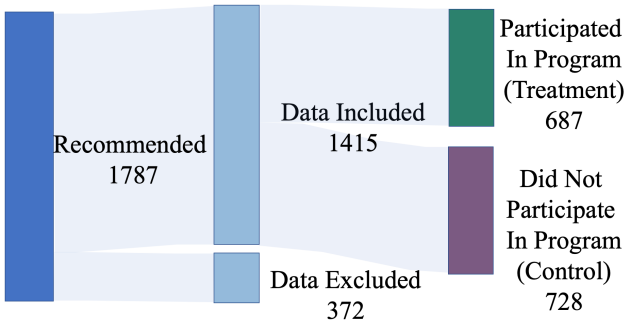

Figure 1: Flow chart illustrating those who were kept and removed from our final dataset, and how many participated in the program.

Impact Evaluation and Analyses

Descriptive Analysis. Detailed descriptive information on the people that we examined is provided in the supplement. Most of these people were Black (73%), male (68%), and 17-30 years old (60%). Most of them were facing misdemeanor or non-violent felony charges (93%). Most of them ultimately had their charges dismissed (82%), but a small percentage of them were convicted (17%) but not sent to prison, and only seven were acquitted. The people who were convicted were sentenced to either jail or community services or both (12%).

Regarding risk levels, the CJRC assessed risk using the *Virginia Pre-trial Risk Assessment Instrument (VPRAI)*. This version of the VPRAI provides risk scores that range from zero to nine, which are based on eight risk factors such as criminal history, employment status, and history of drug abuse (VanNostrand and Rose 2009). Collapsed across treatment and control conditions, these people were low-risk on average (M = 2.5, SD = 1.7). Few individuals had a risk score equal to or greater than 5 (15%). Those in the treatment (M = 2.4, SD = 1.7) and control conditions (M = 2.6, SD = 1.8) had nearly the same average risk score, even with-

out adjusting for other confounders. We assess the accuracy of the VPRAI in the supplementary material.

Of the 687 people who took part in the program, 295 of them had a new criminal charge (43%). Of the 728 people who did not take part in the program, 303 of them had a new criminal charge (42%). These data suggest that the treatment did not significantly reduce new criminal charges. Before making that conclusion though, it is important to minimize confounding differences between the control and treatment conditions. In the next sections, we use traditional and AI-based observational causal inference methods to evaluate that possibility.

Traditional Analysis: Propensity Matching. Here, we evaluate whether the CJRC program had a causal effect on new criminal charges by conducting propensity score matching (Rosenbaum and Rubin 1983). Researchers have used this classic matching method to evaluate pre-trial programs in the past (Sainju et al. 2018). This statistical approach matches people in the treatment and control conditions based on their propensity to be treated, which in this case is the probability to take part in the program. Propensity score matching methods rely on properly identifying the set of important confounders — covariates that influence selection into treatment while still being related to the outcome of interest — and the most common method for that is the reliance on expert knowledge.

The covariates we chose through the help of experts are the following: *gender, race, charge class, number of years at current residence, veteran status, number of prior adult convictions, age, and VPRAI risk score*. While other experts in the field may choose a different set of covariates, there is good reason to think that these covariates are important; many of these covariates are correlated with re-arrests and new criminal activity (Desmarais, Johnson, and Singh 2016; VanNostrand and Rose 2009). Further, in the supplementary material, we show that the VPRAI risk scores had predictive value such that people with higher scores were more likely to be charged with new crimes than people with lower scores. This result replicates recent large-scale evaluations of a variety of risk assessment instruments (Desmarais, Johnson, and Singh 2016) and validates the need to include them as a confounder in this causal analysis. Using these covariates, we computed a propensity score for each individual and then matched people in the treatment and control conditions based on that score.

The proportions of men between the treatment and control conditions became closer after matching; this is because unmatched units were removed from the dataset during the matching process. A similar improvement in distributional balance is seen for women as well. In fact, the difference in proportions for each of our expert-identified covariates was large before matching but was considerably smaller after matching. That is, the distributional balance improved after matching. In many cases, that difference was reduced to zero. A Love plot summarizing this improvement for each of our expert-identified covariates is in the supplement. Having established that the distributional balance of our expert-identified covariates improved after matching, we estimated an average treatment effect of -0.0375, with a 95% CI of $[-0.517, 0.440]$. **We therefore find that the CJRC program neither increased nor decreased new criminal charges, on average, based on this method.**

One glaring problem with propensity score matching is that the matched groups are not interpretable. When we match a set of units with similar propensity, those units may differ substantially on important covariates. For example, we may match units that differ in race, age, and risk level if those units yield a similar propensity score. For this reason, propensity score matching is not a reliable tool for estimating conditional treatment effects. Rather, it can only be used to estimate the *average* treatment effect across individuals. Interpretable machine learning approaches (discussed next) avoid these issues while providing researchers additional capabilities, including the ability to create *conditional* average treatments effects for each individual – not just the average.

Interpretable Machine Learning Analysis

FLAME – Fast Large-scale Almost Matching Exactly (Wang et al. 2021) – and DAME – Dynamic Almost Matching Exactly (Dieng et al. 2019) – are powerful and non-parametric AI-based algorithms that yield fully interpretable matched groups. Both algorithms receive as input a dataset that contains discrete observational covariates, a binary indicator for treatment, and a continuous or discrete outcome column. The treatment indicator in our data is whether or not the individual participated in the pre-trial program, and the outcome column is binary, indicating whether the individual was charged with another crime in the one-year window after their case was disposed.

Outline of FLAME-DAME. The algorithms aim to output a *matched group* for each unit in the matching set. These algorithms use some of the dataset as a *training set* and the rest of the dataset as the *matching set*. (We can also use the full dataset as the matching set if we replicate points from the training set within the matching set, which we do here.) The training set (here, 50% of the dataset) is only used to determine which covariates will be used for matching for each unit. (Different units are matched on different covariates.) During this procedure, matched groups are formed only when there is both a treatment and control unit in each matched group. A process diagram for the FLAME and DAME algorithms is provided in the supplement.

First iteration. The algorithms begin by matching any units (in the matching set) that can be matched exactly to other units on *all* covariates, ensuring both a treatment and a control unit are present in every matched group. This produces a set of matched groups of the highest possible quality, as each matched unit is matched identically to one or more other units on all covariates.

Second iteration: Next, a machine learning method (ridge regression) is applied to the training set to determine an important subset of the covariates – one that predicts the outcome well. Note that the algorithm cannot use all covariates since it has already completed matches on all variables in the first iteration – it must omit at least one covariate. FLAME/-DAME again matches as many units as possible (from the matching set) exactly on the subset of covariates identified by ridge regression.

Subsequent iterations: These two steps are repeated, with the machine learning approach being used to select the new best set of covariates to match on at each iteration.
Stopping conditions: The algorithms stop when either all units have been matched, a pre-determined number of iterations is reached (user's choice), or if the remaining set of covariates no longer is sufficient for the machine learning method to predict the outcome well.

The crucial difference between FLAME and DAME is *how* the best covariate set is selected. FLAME eliminates the least helpful covariate in predicting the outcome at each iteration. It is helpful to think of FLAME as conducting backward selection on the training set. DAME instead eliminates the least important new subset of covariates at each iteration (i.e., as long as that subset of covariates has not been eliminated before), starting with the full set of covariates every time. Thus, DAME is more thorough but takes much longer as it needs to examine each possible subset of covariates.

For our task, we use FLAME for a user-defined number of iterations and then switch to the more computationally-intensive DAME iterations. This allows FLAME to quickly eliminate unhelpful covariates before it transitions to DAME for a finer-grained analysis on a manageable set of covariates. We run the combination of FLAME and then DAME on all 31 covariates, without having any experts identify the 'important' covariates. We ran FLAME for 18 iterations and then switched to DAME, to ensure good quality matched groups within a reasonable runtime. The implementation details are provided in the supplementary material. Also, we have further constrained the matched groups to contain at least 5 treated and 5 control units, rather than 1 of each. The results are not sensitive to this choice, which ensures bigger (higher quality) matched groups.

Results of FLAME-DAME. We inspect the quality and interpretability of the matched groups created using FLAME-DAME, and compare them with the matched groups created using propensity score matching. We demonstrate that FLAME-DAME created better and more interpretable matched groups than propensity score matching. Let us consider Unit 924, a Black male charged with a felony who took part in the pre-trial program. Unit 924 is 18-20 years old, has a risk score of three, and has less than five prior convictions. FLAME-DAME revealed that it produced a 'good' quality matched group for this unit. (We discuss Unit 311 in the supplementary material, which had a 'poor' quality matched group produced, as reported by FLAME. Even FLAME's 'poor' matched groups are better than those of propensity score matching.)

Tables 1 to 5 contain the matched groups for Unit 924. The tables show the expert-identified covariates that were used for matching these units. As a reminder, earlier we identified these covariates as important covariates for matching: *gender, race, charge class, number of years at current residence, veteran status, number of prior adult convictions, age, and VPRAI risk score*. We focus on these covariates in this analysis for that reason. The rest of the covariates are omitted from the tables to show performance on the expert-identified covariates.

Note that the notation (1,5) and (5,5) used in Tables 2 to 5 for propensity score matching refers to the number of treatment and control units in the matched groups, respectively.

Table 1 shows the units FLAME-DAME matched to Unit 924. FLAME-DAME found many units that matched Unit 924 on almost all the expert-identified covariates, and it did so automatically without requiring us to select these covariates beforehand. In fact, one unit from this matched group (i.e., Unit 1255) matched Unit 924 exactly on all these covariates. The remaining units in this group matched Unit 924 on every covariate from this subset except on *age*. We highlight these few mismatches in gray.

Table 2 shows the matched group for Unit 924 provided by propensity score matching. Here, we used the full set of covariates for matching (although we only show the expert-identified covariates in the table). Each matched group was forced to have at least one treatment unit and five control units. Unit 924's matched group, shown in Table 2, did not closely agree on the expert-identified covariates, which suggests that this is a lower quality matched group than that of Table 1. For example, Unit 242 was matched to Unit 924 even though this unit is a 35–39-year-old White woman charged with a misdemeanor.

We then did the same propensity score matching analysis but required at least five treatment and control units per matched group. However, this change did not improve the quality of the matched group. Table 3 shows the matched group for Unit 924 from this analysis. Although the matched group had more units than before (ten vs. five units), these units matched Unit 924 on a small number of covariates. There were many mismatches present, shown in gray.

The previous tables show matched groups that were created by propensity matching on all available covariates. Tables 4 and 5 show the units propensity-matched to Unit 924 when matching only on the expert-identified covariates. Also, the results in Table 4 required at least one treatment and five control units per matched group, whereas the results in Table 5 required at least five treatment and control units per matched group. Restricting the propensity score matching analysis to the expert-identified covariates did not improve the quality of the matched group. Tables 4 and 5 show many units that differ from Unit 924 in important respects (e.g., age, gender, race, and charge class).

Table 1 thus reveals the trustworthiness and interpretability of FLAME-DAME. No matter which propensity score method we used (i.e., requiring one or five treatment units) and which set of covariates we used for matching (i.e., the full set or the subset of expert-identified covariates), FLAME-DAME provided a better quality matched group.

The matched groups provided by FLAME-DAME for Units 924 and 311 (311 is in the supplementary material) were not atypical. Rather, FLAME-DAME provided more cohesive matched groups, on average, than propensity score matching. Figure 2 shows the distribution of the number of expert-identified covariates that matched exactly in each group of units for FLAME-DAME and six different configurations of propensity score matching. More cohesive, higher quality matched groups contain units that match on a large number of these *relevant* covariates. Thus, distri-

Unit	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
924	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
25	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	27 - 29	3
170	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
182	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
289	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	45 - 49	3
322	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
373	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	27 - 29	3
408	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
411	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
471	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	35 - 39	3
512	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	30 - 34	3
591	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	27 - 29	3
744	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
748	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
749	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	27 - 29	3
914	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
985	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
1015	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
1039	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	21 - 23	3
1059	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	40 - 44	3
1130	0	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3
1255	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
1323	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	24 - 27	3

Table 1: FLAME-DAME: a high-quality matched group. A gray cell means there is not an exact match to the query; grayer is worse (less cohesive). FLAME-DAME *found matches on almost all the expert-identified important covariates automatically*.

Unit	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
924	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
1397	0	Female	Black	2 or 3	none	no	0 - 5	27 - 29	2
904	0	Female	Black	H or I	$<$1 year	no	0 - 5	18 - 20	2
242	0	Female	White	A1 or 1	$\geq$ 1 year	no	0 - 5	35 - 39	0
275	0	Female	Black	2 or 3	$<$1 year	no	6 - 10	40 - 44	7
1006	0	Male	Black	H or I	$<$1 year	no	6 - 10	35 - 39	6

Table 2: Propensity Score Matching (1,5) on the full set of covariates. Same query unit as in Table 1, less cohesive matched group than Table 1.

Unit	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
924	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
719	0	Male	Black	2 or 3	$<$1 year	no	0 - 5	24 - 26	5
472	1	Female	Black	A1 or 1	$\geq$ 1 year	no	0 - 5	30 - 34	3
127	1	Female	White	H or I	$<$1 year	no	0 - 5	21 - 23	5
141	0	Male	Black	2 or 3	$\geq$ 1 year	no	0 - 5	21 - 23	1
530	0	Female	Black	A1 or 1	$\geq$ 1 year	no	0 - 5	50 - 54	1
218	0	Male	Hispanic	2 or 3	$<$1 year	no	0 - 5	27 - 29	2
624	0	Male	Black	A1 or 1	$<$1 year	no	0 - 5	30 - 34	1
46	1	Male	Hispanic	H or I	$\geq$ 1 year	no	0 - 5	24 - 26	2
488	1	Male	Black	A-G	$\geq$ 1 year	no	16 - 20	40 - 44	5
188	1	Male	Hispanic	A1 or 1	$<$1 year	no	0 - 5	35 - 39	1

Table 3: Propensity Score Matching (5,5) on the full set of covariates. Same query unit as in Table 1, less cohesive matched group than Table 1.

butions toward the right depict better matched groups than distributions toward the left. The matched groups provided by FLAME-DAME is represented by the light-blue distribution, and the mean of that distribution falls further to the right than the mean of any other distribution. This shows that FLAME-DAME provided more cohesive and more interpretable matched groups than the propensity score matching methods.

Unit	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
924	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
82	0	Male	Other	A1 or 1	<1 year	no	0 - 5	30 - 34	1
1322	0	Female	Black	H or I	<1 year	no	0 - 5	21 - 23	4
1226	0	Male	Black	H or I	<1 year	no	0 - 5	18 - 20	4
157	0	Female	Black	2 or 3	$\geq$ 1 year	no	0 - 5	21 - 23	5
451	0	Male	Black	H or I	<1 year	no	0 - 5	21 - 23	3

Table 4: Propensity Score Matching (1,5) on the expert-identified covariates. Same query unit as in Table 1, less cohesive matched group than Table 1.

Unit	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
924	1	Male	Black	H or I	$\geq$ 1 year	no	0 - 5	18 - 20	3
46	1	Male	Hispanic	H or I	$\geq$ 1 year	no	0 - 5	24 - 26	2
680	1	Female	White	A1 or 1	$\geq$ 1 year	no	0 - 5	27 - 29	3
642	1	Male	Other	A1 or 1	<1 year	no	0 - 5	40 - 44	5
79	0	Male	Black	H or I	<1 year	no	20+	55 - 59	4
236	0	Female	Black	A1 or 1	<1 year	no	0 - 5	24 - 26	1
605	0	Male	Black	2 or 3	$\geq$ 1 year	no	16 - 20	55 - 59	2
235	1	Female	Black	A1 or 1	<1 year	no	6 - 10	30 - 34	7
434	0	Male	Hispanic	A1 or 1	$\geq$ 1 year	no	0 - 5	35 - 39	0
495	1	Female	White	2 or 3	$\geq$ 1 year	no	0 - 5	40 - 44	1
200	0	Female	Black	A1 or 1	$\geq$ 1 year	no	0 - 5	30 - 34	0

Table 5: Propensity Score Matching (5,5) on the expert-identified covariates. Same query unit as in Table 1, less cohesive matched group than Table 1.

Additionally, we constrained FLAME-DAME to have at least five treatment and five control units in each matched group. However, we relaxed that constraint for some configurations of propensity score matching (such as those represented by the yellow and orange colored distributions). These configurations permitted smaller matched groups to form, making it easier to match on a large number of expert-identified covariates. However, since smaller matched groups tend to be less reliable, our goal is to have a higher number of matching covariates while also maintaining larger matched groups. FLAME-DAME achieves this goal better than propensity score matching because it provided higher quality matched groups while also maintaining larger matched groups (i.e., more units in each group).

These results strongly suggest that the matched groups provided by FLAME-DAME are trustworthy, and are likely to be much more trustworthy than the matched groups provided by the propensity score matching.

Having established the trustworthiness of our matched groups, we now test whether the treatment had a significant effect on our outcome of interest. Using FLAME-DAME, we obtain an estimated average treatment effect of 1.8% with the 95% confidence interval including zero. **Thus, this more sophisticated and rigorous machine-learning-based approach also found no evidence that the pre-trial program significantly increased or decreased the probability of a new criminal charge.** This null result solidifies the null result found using propensity score matching. Further, there were no identifiable subgroups for which a non-zero treatment effect was significant; that is, the program did not appear to benefit any known subgroup of individuals.

Stability of Null Result

While we have obtained null results from the traditional method (propensity score matching) and the modern method (FLAME-DAME), the null results persist even when we consider a variety of black box methods for observational causal inference. Table 6 shows ATE results from Causal BART (Chipman et al. 2010; Hill, Reiter, and Zanutto 2004) (here implemented as difference of BARTs on the treated group and the control group), Genetic Matching (Diamond and Sekhon 2013), and Mahalanobis distance matching (Rubin 1980), all of which indicate that the treatment effect from the program is small and not significantly different from zero.

Method	ATE	Interpretable
FLAME-DAME	0.0198 (null)	Yes
Propensity Score	-0.0375 (null)	No
Mahalanobis Distance	-0.0322 (null)	No
GenMatch	-0.0315 (null)	No
Causal BART	0.0143 (null)	No

Table 6: Black box machine learning based methods for causal inference all yield a small average treatment effect (ATE) that is not significantly different from 0.

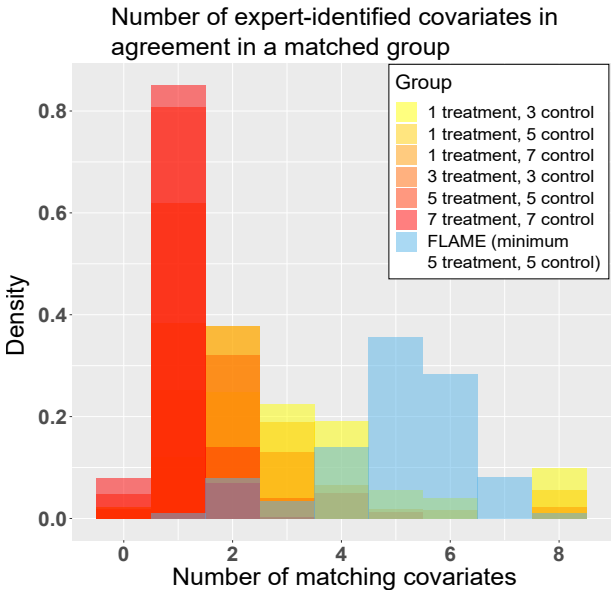

Figure 2: Histograms for the number of expert-identified covariates in agreement in a matched group, for FLAME-DAME along with 6 different configurations of propensity score matching. The legend denotes the size of the matched group, with the propensity score matching configurations outputting matched groups of fixed size with the shown ratio of treatment units to control units. On the whole, the number of expert-identified covariates in agreement in a matched group for FLAME-DAME are higher than the selected propensity score matching configurations. Smaller matched groups tend to be less reliable, so we desire a higher number of matching covariates (towards the right of the plot) with larger matched groups.

Discussion

As discussed earlier, approximately five-hundred thousand people are detained pre-trial each day in the United States. The effect of pre-trial policies on such a large number of people raises important policy, legal, and constitutional issues (Heaton, Mayson, and Stevenson 2017). We know that many people are detained simply because they cannot pay cash bail (Minton and Zeng 2021), and that pre-trial detention has criminogenic effects, which lead to increased recidivism and other economic and social harms (Dobbie, Goldin, and Yang 2018; Heaton, Mayson, and Stevenson 2017; Leslie and Pope 2017; Lowenkamp and VanNostrand 2013). Jurisdictions have engaged in significant efforts to rethink their pre-trial policies in recent years (Garrett 2022). As part of these efforts, alternatives to pre-trial detention have become an important area for policy and for research.

While community-based supervision may be better than pre-trial detention, there is growing concern that these programs, and the burdens that they impose, could still be harmful, or at the least, not a good use of taxpayer funds. Bechtel et al. (2017) strongly emphasized that the field of pre-trial research is in need of much improvement, especially regarding the statistical tools used for program evaluation. *We addressed this important problem in this article by introducing a set of new interpretable machine learning matching techniques to the field of pre-trial research.* These techniques can rigorously evaluate the effectiveness of pre-trial programs. In comparison to older matching approaches such as propensity score matching, these new techniques allow researchers to examine average treatment effects and conditional treatment effects without needing to rely on subjective assessments of covariate importance (among other benefits).

Consistent with the recent meta-analyses by Bechtel et al. (2017), we find that the CJRC pre-trial program located in Durham, NC neither reduced nor increased new criminal charges. That is, we found an average treatment effect that was approximately zero. Further, we found no measurable effects for any subgroups within this population. That is, there was no evidence that the program benefited some smaller group of people, in terms of reducing their probability of being charged with a new crime.

Based on these and prior results, it seems that pre-trial programs are generally ineffective at reducing new criminal charges. If this effect replicates, the criminal-legal system needs to focus resources. The simplest option is to release low-risk individuals back into the community while they await trial without subjecting them to any restrictions or conditions. This type of pre-trial release was the norm in the United States for many years until relatively recently. If more burdensome pre-trial programs do not have substantial benefits, then simple release alternatives should be more carefully considered. Reflecting on the low-risk population we examined, it is not clear what the public safety goals were originally in these cases. Eighty-two percent of these low-level criminal cases were eventually dismissed. Subjecting people to supervision for cases that will ultimately be dismissed may not accomplish useful public safety goals, and the outcomes in these cases raise questions about whether these charges were sufficiently warranted in the first place.

The interpretable machine learning techniques used here have the potential for use in a wide variety of areas in which randomized assignment is not possible, ethical, or feasible. In addition to pre-trial programs, these techniques can be used to evaluate other social programs such as felony diversion programs. These techniques are not suitable for program evaluation if there is an insufficient amount of data or error-prone data. We are grateful to the CJRC for their meticulous care in collecting data. We encourage other agencies to do the same.

These data pose a larger challenge: while risk assessment instruments, such as the VPRAI used as part of this pre-trial program, may be quite predictive, that risk tool was used to place people in a program that itself did not reduce the relevant risks. We need empirically validated ways to reduce the risks that risk assessments can identify. Improved and interpretable program evaluation tools can address that more fundamental challenge facing so many social programs.

Acknowledgments

We acknowledge support from the National Science Foundation under grant NSF IIS-2147061, and from the Wilson Center for Science and Justice at Duke Law School.

References

Austin, J.; Krisberg, B.; and Listky, P. 1985. The effectiveness of supervised pretrial release. *Crime and Delinquency*, 31(4): 519–537.

Bechtel, K.; Holsinger, A. M.; Lowenkamp, C. T.; and Warren, M. J. 2017. A meta-analytic review of pretrial research: Risk assessment, bond type, and interventions. *American Journal of Criminal Justice*, 42(2): 443–467.

Cadigan, T.; and Lowenkamp, C. 2011. Implementing Risk Assessment in the Federal Pretrial Services System. *Federal Probation*, 75(2).

Chipman, H. A.; George, E. I.; McCulloch, R. E.; et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1): 266–298.

Clark, J. W.; and Henry, D. A. 2003. *Pretrial services programming at the start of the 21st century: A survey of pretrial services programs*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Clarke, S. H. 1988. Pretrial release, concepts, issues and strategies for improvement. *Research in Corrections*, 1(3): 1–40.

Cohen, T. H. 2012. *Pretrial release and misconduct in federal district courts, 2008-2010*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Cooprider, K.; and Kerby, J. 1990. Pretrial release, concepts, issues and strategies for improvement. *Federal Probation*, 54.

Desmarais, S. L.; Johnson, K. L.; and Singh, J. P. 2016. Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services*, 13(3): 206.

Diamond, A.; and Sekhon, J. S. 2013. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics*, 95(3): 932–945.

Dieng, A.; Liu, Y.; Roy, S.; Rudin, C.; and Volfovsky, A. 2019. Interpretable almost-exact matching for causal inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2445–2453. PMLR.

Dobbie, W.; Goldin, J.; and Yang, C. 2018. The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review*, 108(2): 201–240.

Garrett, B. 2022. Models of Bail Reform. *Florida Law Review*.

Goldkamp, J. S.; and White, M. D. 2006. Restoring accountability in pretrial release: The Philadelphia pretrial release supervision experiments. *Journal of Experimental Criminology*, 2(2): 143–181.

Heaton, P.; Mayson, S.; and Stevenson, M. 2017. The Downstream Consequences of Misdemeanor Pretrial Detention. *Stanford Law Review*, 69(3): 711–717.

Hill, J. L.; Reiter, J. P.; and Zanutto, E. L. 2004. A comparison of experimental and observational data analyses. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, 49–60.

Iacus, S. M.; King, G.; and Porro, G. 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1): 1–24.

Lanners, Q.; Parikh, H.; Volfovsky, A.; Rudin, C.; and Page, D. 2023. Variable Importance Matching for Causal Inference. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*.

Leslie, E.; and Pope, N. 2017. The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from NYC Arraignments. *Journal of Law and Economics*, 60(3): 529–547.

Levin, D. J. 2007. *Examining the Efficacy of Pretrial Release Conditions, Sanctions and Screening with State Court Processing Statistics Dataseries*. PJI, Pretrial Justice Institute.

Lowenkamp, C.; VanNostrand, M.; and Holsinger, A. 2013. The Hidden Costs of Pretrial Detention. *The Laura and John Arnold Foundation*, 1–32.

Lowenkamp, C. T.; and VanNostrand, M. 2013. *Exploring the impact of supervision on pretrial outcomes*. Laura and John Arnold Foundation Houston, TX.

Mahoney, B.; Beaudin, B.; Carver, J.; Ryan, D.; and Hoffman, R. 2001. Pretrial Services Programs: Responsibilities and Potential. *Issues and Practices in Criminal Justice*, 1 – 122.

Mamalian, C. A.; et al. 2011. State of the science of pretrial risk assessment. *Pretrial Justice Institute*.

Minton, T.; and Zeng, Z. 2021. Jail Inmates in 2020 – Statistical Tables. *U.S. Department of Justice Office of Justice Programs Bureau of Justice Statistics*, 1 – 27.

Parikh, H.; Rudin, C.; and Volfovsky, A. 2022. MALTS: Matching After Learning to Stretch. *Journal of Machine Learning Research*, 23(240): 1–42.

Parmer, G.; and Merritt, A. L. 2019. Durham county criminal justice resource center provides compassionate and innovative care for justice-involved community members. *North Carolina Medical Journal*, 80(6): 369–370.

Phillips, M. 2012. A Decade of Bail Research in New York City. *U.S. Department of Justice Office of Justice Programs*, 1–116.

Robinson, C.; VanBenschoten, S.; Alexander, M.; and Lowenkamp, C. 2011. A Random (Almost) Study of Staff Training Aimed at Reducing Re-arrest (STARR): Reducing Recidivism through Intentional Design. *Federal Probation*, 75(2).

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 1980. Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics*, 36(2): 293–298.

Sainju, K.; Fahy, S.; Baggaley, K.; Baker, A.; Minassian, T.; and Filippelli, V. 2018. Electronic Monitoring for Pretrial Release: Assessing the Impact. *Federal Probation*, 82(3): 3–10.

Stevenson, M. 2018. Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes. *Journal of Law, Economics, and Organization*, 34: 511–532.

Van Der Laan, M. J.; and Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

VanNostrand, M.; and Keebler, G. 2009. Pretrial risk assessment in the federal court. *Fed. Probation*, 73: 3.

VanNostrand, M.; and Rose, K. 2009. The Virginia Pretrial Risk Assessment Instrument. *Pretrial Risk Assessment in Virginia*, 1–27.

VanNostrand, M.; Rose, K. J.; and Weibrecht, K. 2016. State of the Science of Pretrial Release Recommendations and Supervision (2011). *Pretrial Justice Institute*.

Wager, S.; and Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wang, T.; Morucci, M.; Awan, M. U.; Liu, Y.; Roy, S.; Rudin, C.; and Volfovsky, A. 2021. FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference. *J. Mach. Learn. Res.*, 22: 31–1.

Dataset Overview

Table 3 provides an overview of our dataset.

		Frequency	Percentage
Condition	Control	728	51%
	Treatment	687	49%
Race	Black	1082	76%
	Hispanic	129	9%
	Native American	2	0%
	Other	16	1%
	White	185	13%
Gender	Female	466	33%
	Male	948	67%
	Undefined	1	0%
Age	17-20	305	22%
	21-25	283	20%
	26-30	259	18%
	31-35	157	11%
	36-40	125	9%
	41-45	78	6%
	46-50	74	5%
	51-55	61	4%
	56-60	39	3%
	61+	34	2%
Crime	Assaultive Misdem.	679	44%
	Non Assaultive Misdem.	279	18%
	Non Violent Felony	363	24%
	Traffic	35	2%
	Violent Felony	59	4%
Case Outcome	Charges dismissed	1164	82%
	Guilty to Lesser Degree	7	0%
	Guilty	237	17%
	Not Guilty	7	0%
Sentence Outcome	Jail / Commun. Serv.	166	12%
	No Sentence	1249	88%
New Criminal Charge	No	817	58%
	Yes	598	42%

Table 3: Descriptive statistics for the dataset.

Reproducibility

This research project used three separate datasets including the data from the Durham County Jail, the Criminal Justice Resource Center (CJRC), and the North Carolina Administrative Office of the Courts (AOC) Automated Criminal Infraction System (ACIS). The CJRC and the Durham County Jail data are not publicly available. The ACIS extract is publicly available and can be found on the AOC website (AOC 2022). The code used to process and analyze these data is available upon request. The FLAME-DAME python package is also publicly available and can be found online with full documentation and examples (FLAME and DAME packages 2019). Parameter settings for FLAME-DAME and other packages can be found further ahead in the supplement.

Data Collection

We importantly note that noise and inconsistency in the CJRC, Durham County jail, and ACIS data arose for many reasons, including the fact that data entries were made by different human operators at different times. For example, there are missing middle or first names, spelling mistakes in names, and other typographical errors. At times, digits in birthdates were swapped. Unfortunately, there was no way to link individuals across the three datasets easily because there was no unique personal ID. For these reasons, we had to contend with these errors and missing information. We linked people using personal identifying information such as name and date of birth. Fortunately, we have the first and last name of every defendant, we have two examples of how the defendant name has been entered (`CLNT_LAST_NAME`, `CLNT_FIRST_NAME` columns for the CJRC part, and `CRRNAM` column for the ACIS part), and the same for how the date of birth has been entered (`CLNT_DATE_OF_BIRTH` for the CJRC part, and `CRRDOB` for the ACIS part). We utilized both exemplars in our searching process used to obtain the repeat arrest (as defined above) data. Here we describe this process:

- First, we preprocessed the name columns, removing extraneous characters such as extra spacing and rearranging suffixes such as Jr., Sr., I, II, III, etc. to ensure the ordering: `lastname, middlename, firstname, suffix`.
- Then we looped through the *full*, up-to-date ACIS dataset, extracting people whose last name (similarly preprocessed) is an exact match for either `CLNT_LAST_NAME` or `lastname` extracted from `CRRNAM`, and whose date of birth is an exact match for either `CLNT_DATE_OF_BIRTH` or `CRRDOB` (special characters such as apostrophe, comma, full stop, hyphen, extra spacing, etc. notwithstanding).
- We further refined the set of people extracted as above, by keeping only those for whom *at least one* of their (again, similarly preprocessed and comma-separated) middle or first names (each of which may itself be composed of multiple words) is an exact match for any one of the comma-separated words in either `CLNT_FIRST_NAME` or `middlename or firstname` extracted from `CRRNAM` (of course, this excludes the suffixes or other characters).
- Next, we kept only those entries in this set whose date of indictment, charge levied, or arrest is within 1 year of the CJRC case disposal date, which as discussed, is our measure of repeat arrest.
- Finally, we refined this dataset of repeat arrest cases by searching through the charged offense codes, and removing all low-level traffic offenses since these are not outcomes that the CJRC program is intended to reduce.

We also searched through the dataset to obtain the punishment outcomes for the CJRC trial, i.e., 3 indicators denoting

'incarceration for a time,' 'no incarceration,' and 'released but require community service,' and removed all individuals who were incarcerated for any period of time to ensure that the 1-year post-trial window remains fair for everyone (i.e., the defendant is out of prison, and the possibility of future arrest exists).

Predictive Value of the VPRAI

The primary goal of this study is to determine whether the pre-trial program operated by the CJRC increased or decreased new criminal charges, which is a causal question, not a predictive question. We introduce state-of-the-art interpretable machine learning techniques for observational causal inference to do so. However, we take a moment here to analyze the predictive value of the Virginia Pre-trial Risk Assessment Instrument (VPRAI). The CJRC used the revised version of the VPRAI to calculate a risk score. This version of the VPRAI used eight risk factors such as primary charge type, criminal history, and length at current residence to provide risk scores that ranged from 0 to 9. More information about this instrument is detailed by VanNostrand and Rose (2009). We believe researchers may be interested in this analysis given the substantial focus on improving risk assessment instruments used by the criminal-legal system (e.g., Zeng, Ustun, and Rudin (2017); Wang et al. (2022)).

We first analyzed the VPRAI risk scores by conducting receiver operating characteristic (ROC) analysis. The area underneath (AUC) the ROC is arguably the most important measure of classification performance. Given two individuals, one who was charged with a new crime and one who was not, the AUC is the fraction of such pairs that were ranked correctly by their VPRAI scores. The ROC curve for the VPRAI scores is shown in Figure 3A. We find an AUC equal to 0.63, which means that the VPRAI correctly ranked individuals with new criminal charges above those without new criminal charges 63% of the time. This performance metric replicates a large-scale meta-analysis that investigated a variety of risk assessment instruments (Desmarais, Johnson, and Singh 2016). Desmarais, Johnson, and Singh (2016) often found that pre-trial risk assessment instruments achieve an AUC between 0.6 and 0.7.

Another measure of interest to researchers is the correspondence between risk score and the probability of a new criminal charge (i.e., calibration). This analysis differs from the ROC analysis mentioned previously. This analysis tells us the probability that someone has a new criminal charge given their risk score. Figure 3B shows that the probability of a new criminal charge increases as risk increases. This means that individuals who receive a higher risk score are more likely to be charged with new crimes than those who receive a lower risk score. The size of the points in Figure 3B reflects the number of people who received that risk score. The majority of the people eligible for the pre-trial program were low-risk, which is why those data points are large. Very few people with a high-risk score were eligible to take part in the pre-trial program, which is why those data points are small. Together, the results shown in Figure 3A and 3B suggest that the VPRAI has predictive value.

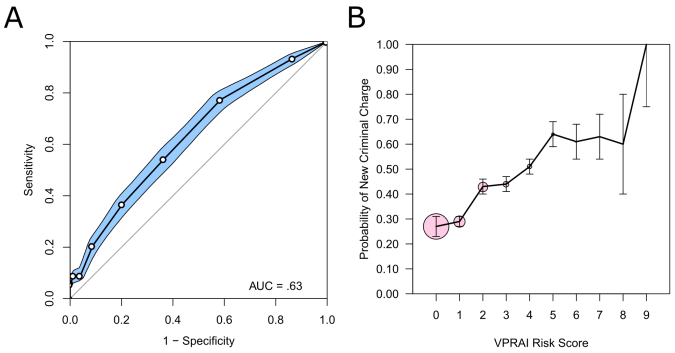

Figure 3: ROC = receiver operating characteristic; AUC = area underneath the curve. Bars represent 95% confidence intervals.

Gender Balancing

Figure 4 shows the improvement in distributional balance for Gender after propensity matching.

Consider the proportion of men in the control condition as well as the proportion of men in the treatment condition. As shown in Figure 4, the difference in the proportion of men between the control and treatment conditions is moderate prior to matching (i.e., for the unadjusted sample). However, that difference is reduced after matching (i.e., for the adjusted sample) because any unmatched units were removed from the dataset. Now, the control and treatment conditions have nearly the same proportion of men.

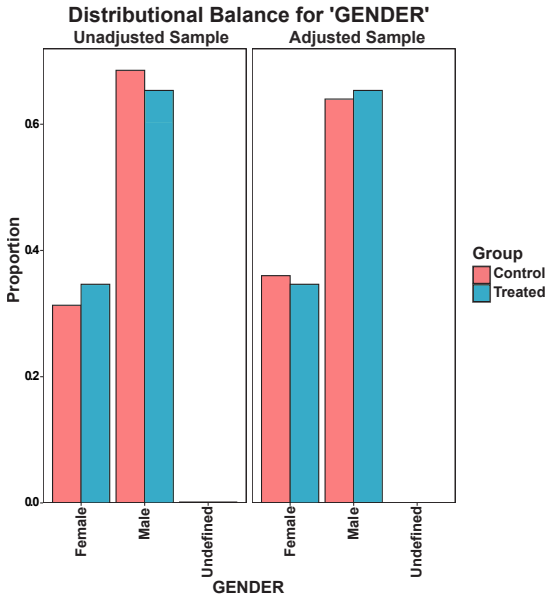

Figure 4: This figure shows the improvement in distributional balance after matching for Gender.

Covariate Balancing

Figure 5 summarizes the distributional balance of the 8 expert-identified covariates before and after matching. It is

clear that the balance was quite poor before matching, but matching (and eliminating data that cannot be matched) improved the balance on all covariates.

Implementation Details

Propensity Score Matching Settings

We used the R libraries `MatchIt`[1] and `Zelig`[2] to implement propensity score matching.
For `MatchIt`, we modified the following hyperparameters, keeping the rest at their default values:-

- *method* = "nearest"
- *distance* = "glm"
- *ratio* = 5
- *replace* = TRUE

For `Zelig`, which was used to compute the ATE, we set the hyperparameter *model* to "ls".

FLAME-DAME Algorithm

Figure 6 shows a flow chart of the FLAME and DAME algorithms over iterations. First, the data are split into a training set and a matching set. The training set will only be used to determine which variables will be used for matching. In the first iteration, all units in the matching set that can be exactly matched on all variables are matched. The matched groups from this operation are shown in the leftmost dark purple box labeled "Matched Groups." Then, the training set is used to determine which variables to use for matching at the second iteration. It cannot use all variables since it has already completed matches on all variables in the first iteration – it must omit at least one variable. After variables are chosen, the matched groups are formed (again in dark purple). At the third iteration, again, the training set is used to determine variables to match on, and there are several iterations like this. In the last iteration, if there are units that cannot be matched on a set of predictive variables, we do not match them (these units are represented by the box labeled "No Match").

Finally, the matched groups are gathered from all iterations. From each matched group, we estimate a conditional average treatment effect for every unit in the dataset.

FLAME-DAME Settings

We used the `dame-flame`[3] Python library to implement FLAME-DAME. Note: We modified this implementation so the algorithms output matched groups with at least 5 control and 5 treated units.
Treating the full dataset as the matching set and 50% of it as the training set, and using the "FLAME" function to run the hybrid FLAME-DAME method, we modified the following hyperparamters, keeping the rest at their default values:

- *replace* = False
- *adaptive_weights* = "RidgeCV"
- *missing_indicator* = -1
- *missing_data_replace* = 2
- *missing_holdout_replace* = 1
- *pre_dame* = 18

Causal BART Implementation

We used the `dbarts`[4] R package to implement Causal BART for binary outcomes by taking a difference of BARTs implemented on the treated group and the control group. We used the "bart" function from the package, keeping all settings at their default values.

Other Black box Approaches

Like Propensity Score Matching, for Mahalanobis Distance Matching and Genetic Matching, we use the `MatchIt` package, keeping only the expert-identified covariates in the dataset, and keeping all settings (except *ratio*, set to 5, and *replace*, set to TRUE) at their default values. For computing the ATE, we again use `Zelig`, setting *model* to "ls".

A 'poor' quality matched group

In the main text we discussed Unit 924 who had a good quality matched group from FLAME-DAME, whereas now we show Unit 311 who FLAME-DAME was unable to find a good matched group for. That is, there was no other unit similar to him on important covariates.

Units 924 and Unit 311 are both Black males charged with H or I felonies who took part in the pre-trial program. Unit 311 is 35-39 years old, has a risk score of four, and has between 6 and 10 prior convictions. We will show the expert-identified covariates that were used for matching Unit 311. As a reminder, we identified these covariates as important covariates for matching: *gender, race, charge class, number of years at current residence, veteran status, number of prior adult convictions, age, and VPRAI risk score*.

Table 4 shows the matched groups resulting from the three different matching algorithms for a treated unit (Unit 311) from a 'poor' matched group (where only a few covariates were able to be matched exactly). Notation (1,5) and (5,5) for propensity score matching refers to the number of treatment and control units in the matched groups, respectively. A red highlighted cell means there is not an exact match between a query unit's covariate and that of a unit in its matched group; more red is worse (less cohesive). The table reveals the trustworthiness and interpretability of the matches produced by FLAME-DAME, which even for a poor quality matched group, was still able to find more (and at least half) of the expert-identified important covariates automatically than the other two methods. (The rest of the covariates are omitted from the table to show the methods' performance on the expert-identified covariates).

Table 4 shows the matched groups for Unit 311. FLAME-DAME produced a 'poor' quality matched group for this unit, which is shown in Table 4(a). Regardless, FLAME-DAME still provided a more cohesive matched group for

[1] https://cran.r-project.org/web/packages/MatchIt/index.html
[2] https://cran.r-project.org/web/packages/Zelig/index.html
[3] https://github.com/almost-matching-exactly/DAME-FLAME-Python-Package
[4] https://cran.r-project.org/web/packages/dbarts/dbarts.pdf

Table 4: FLAME-DAME's 'low-quality' matched group is more cohesive than all variations of propensity score matching.

Unit ID	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
Query: 311	1	Male	Black	H or I	≥ 1 year	no	6 - 10	35 - 39	4
3	0	Male	Black	H or I	≥ 1 year	no	16 - 20	30 - 34	4
70	0	Male	Black	A1 or 1	≥ 1 year	no	11 - 15	60 - 64	1
75	0	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	30 - 34	2
147	0	Male	Black	A1 or 1	≥ 1 year	no	20+	55 - 59	5
225	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	55 - 59	2
248	1	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	24 - 26	4
279	0	Male	Black	2 or 3	≥ 1 year	no	11 - 15	45 - 49	1
296	0	Male	Black	A1 or 1	≥ 1 year	no	16 - 20	40 - 44	1
300	0	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	27 - 29	1
330	1	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	21 - 23	3
370	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	45 - 49	4
400	1	Male	Black	2 or 3	≥ 1 year	no	20+	65 - 69	1
402	0	Male	Black	A1 or 1	≥ 1 year	no	20+	45 - 49	3
488	1	Male	Black	A-G	≥ 1 year	no	16 - 20	40 - 44	5
616	0	Male	Black	H or I	≥ 1 year	no	11 - 15	40 - 44	3
673	1	Male	Black	2 or 3	≥ 1 year	no	16 - 20	35 - 39	5
768	1	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	30 - 34	4
780	0	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	35 - 39	3
817	0	Male	Black	H or I	≥ 1 year	no	16 - 20	40 - 44	5
939	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	45 - 49	0
976	1	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	35 - 39	5
1056	1	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	24 - 26	1
1207	1	Male	Black	2 or 3	≥ 1 year	no	6 - 10	30 - 34	3
1269	1	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	35 - 39	5
1370	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	30 - 34	1

(a) FLAME-DAME: a 'low-quality' matched group, as assessed by the FLAME-DAME algorithm.

Unit ID	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
Query: 311	1	Male	Black	H or I	≥ 1 year	no	6 - 10	35 - 39	4
905	0	Male	White	A1 or 1	≥ 1 year	no	0 - 5	30 - 34	1
551	0	Male	Black	H or I	≥ 1 year	no	0 - 5	21 - 23	2
593	0	Male	Black	H or I	<1 year	no	0 - 5	30 - 34	5
172	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	30 - 34	4
766	0	Male	Black	A1 or 1	<1 year	no	0 - 5	35 - 39	2

(b) Propensity Score Matching (1,5) on the full set of covariates

Unit ID	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
Query: 311	1	Male	Black	H or I	≥ 1 year	no	6 - 10	35 - 39	4
473	0	Male	Black	A1 or 1	<1 year	no	0 - 5	40 - 44	3
291	1	Male	White	A1 or 1	≥ 1 year	no	0 - 5	40 - 44	0
312	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	40 - 44	2
87	0	Male	Black	A1 or 1	≥ 1 year	no	20+	40 - 44	2
395	1	Male	Black	A1 or 1	<1 year	no	0 - 5	30 - 34	1
452	1	Male	Black	H or I	<1 year	no	0 - 5	35 - 39	8
147	0	Male	Black	A1 or 1	≥ 1 year	no	20+	55 - 59	5
405	1	Female	Black	A1 or 1	<1 year	no	11 - 15	40 - 44	4
129	1	Male	White	2 or 3	<1 year	no	0 - 5	24 - 26	2
587	0	Male	Black	A1 or 1	≥ 1 year	no	6 - 10	30 - 34	4

(c) Propensity Score Matching (5,5) on the full set of covariates

Unit ID	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
Query: 311	1	Male	Black	H or I	≥ 1 year	no	6 - 10	35 - 39	4
340	0	Male	Black	A1 or 1	<1 year	no	11 - 15	30 - 34	5
709	0	Male	Black	A1 or 1	<1 year	no	0 - 5	35 - 39	3
997	0	Male	Black	A1 or 1	<1 year	no	0 - 5	35 - 39	3
213	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	35 - 39	2
891	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	35 - 39	2

(d) Propensity Score Matching (1,5) on the expert-identified covariates only

Unit ID	treated	GENDER	RACE	CLASS CHARGE	YEARS AT RESIDENCE	VETERAN	PRIOR ADULT CONVICTIONS	AGE	VPRAI SCORE
Query: 311	1	Male	Black	H or I	≥ 1 year	no	6 - 10	35 - 39	4
176	0	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	40 - 44	2
363	0	Female	Black	A1 or 1	<1 year	no	0 - 5	40 - 44	1
524	1	Male	Black	A1 or 1	≥ 1 year	no	0 - 5	30 - 34	0
111	1	Male	Black	H or I	<1 year	no	0 - 5	18 - 20	3
461	0	Male	Black	A1 or 1	<1 year	no	0 - 5	30 - 34	1
25	0	Male	Black	H or I	≥ 1 year	no	0 - 5	27 - 29	3
244	1	Female	Black	2 or 3	≥ 1 year	no	0 - 5	65 - 69	0
608	1	Female	Black	H or I	≥ 1 year	no	16 - 20	50 - 54	4
53	0	Male	Black	2 or 3	<1 year	no	11 - 15	50 - 54	3
351	1	Male	Black	2 or 3	≥ 1 year	no	0 - 5	18 - 20	0

(e) Propensity Score Matching (5,5) on the expert-identified covariates only

this unit than the matched groups provided by propensity score matching, shown in Tables 4(b), (c), (d), and (e). Again, we matched Unit 311 in a variety of ways. The units shown in Tables 4(b) and (c) were matched to Unit 311 on all covariates requiring at least five control units and either one or five treatment units per group, respectively. The units shown in Tables 4(d) and (e) were matched to Unit 311 on the expert-identified covariates and required at least five control units and either one or five treatment units per group, respectively. No matter which propensity score method we used (i.e., requiring one or five treatment units) and which set of covariates we used for matching (i.e., the full set or the subset of expert-identified covariates), FLAME-DAME provided a better quality matched group for these two units.

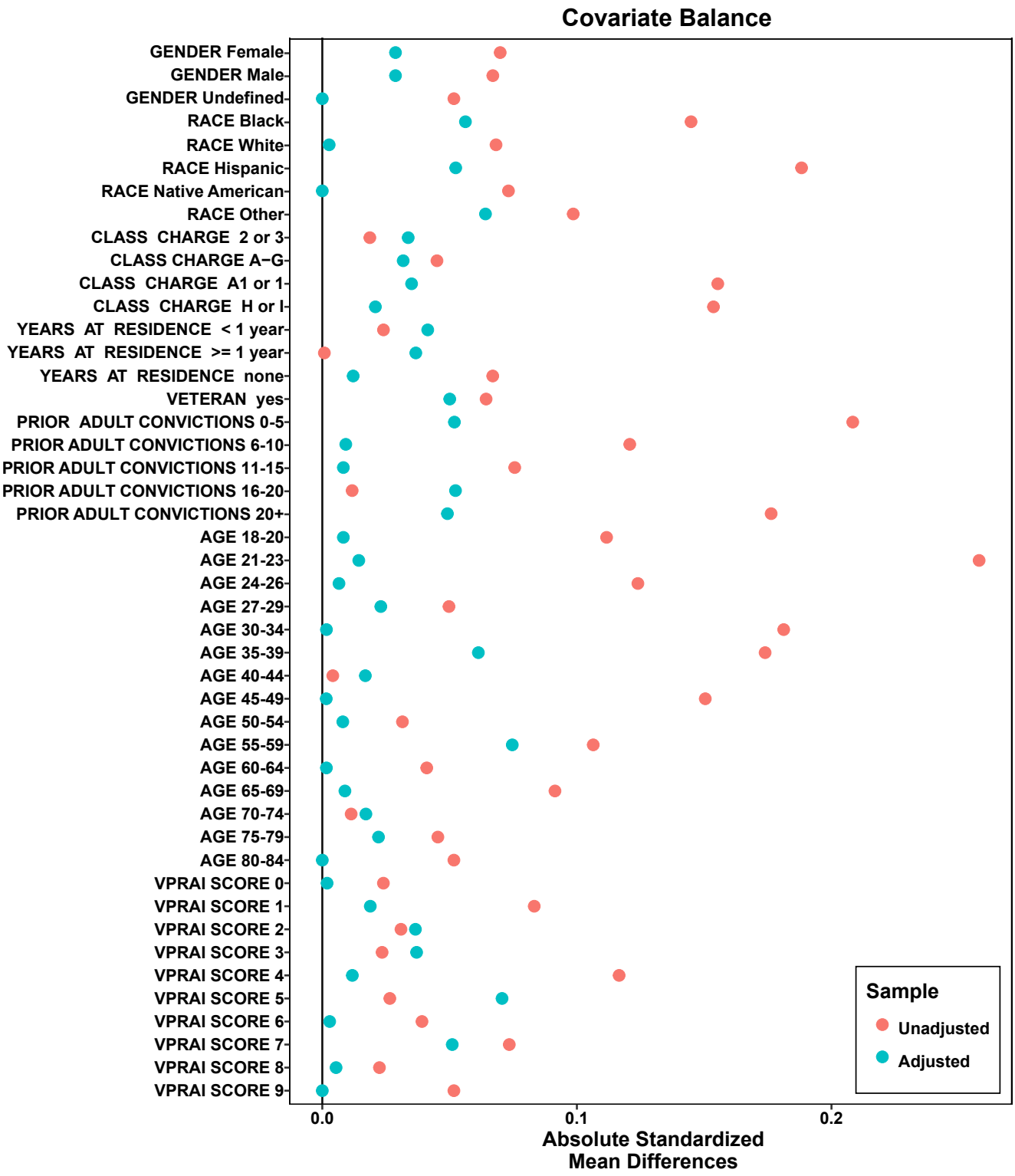

Figure 5: This plot summarizes the distributional balance of the 8 expert-identified covariates before and after matching. It is clear that the balance was quite poor before matching, but matching (and eliminating data that cannot be matched) improved the balance on all covariates.

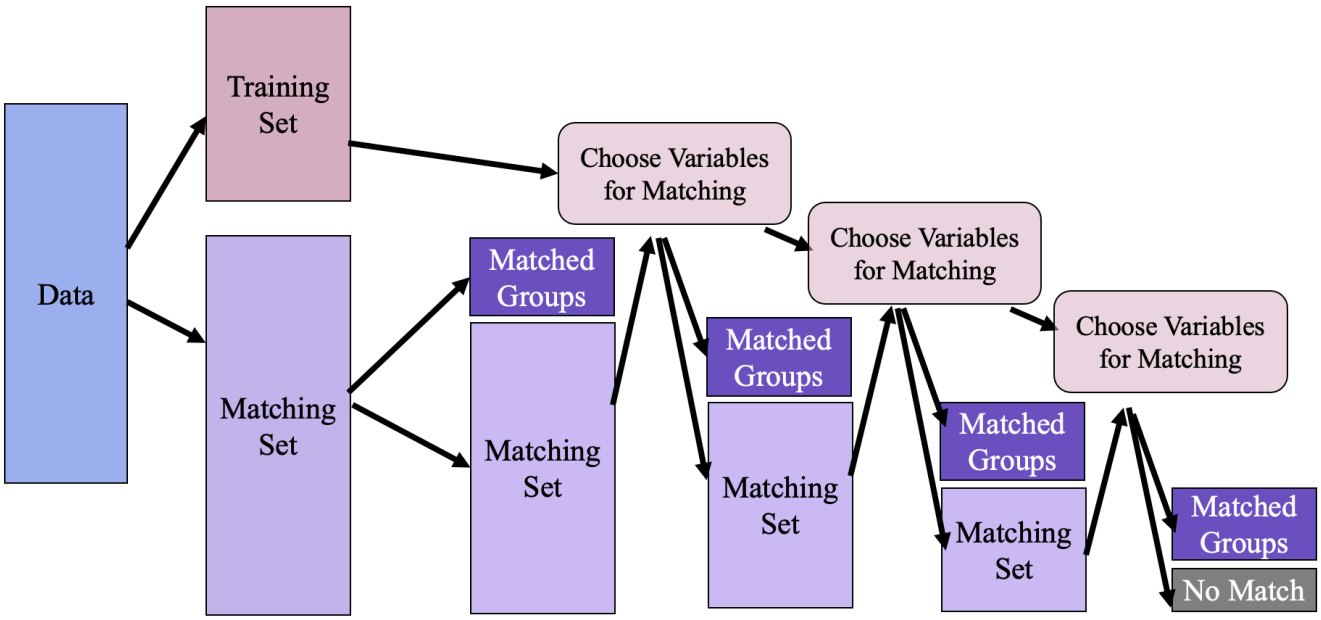

Figure 6: FLAME Flow Chart. Iterations proceed from left to right. The training set is only used to determine variables for matching. At each iteration, we create as many matched groups as possible using the chosen variables. Thus, we need to choose a different set of variables at each iteration. The goal is to get as many high-quality matched groups as possible. Note that every set of variables chosen for every iteration is capable of forming a good predictive model on the training set; thus, every set of variables is good for matching.