Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1053

Unpacking the Impacts of a Youth Behavioral Health Intervention: Experimental Evidence from Chicago

Nour Abdul-Razzak
University of Chicago

Kelly Hallberg
University of Chicago

Racial disparities in violence exposure and criminal justice contact are a subject of growing policy and public concern. We conduct a large-scale, randomized controlled trial of a six-month behavioral health intervention combining intensive mentoring and group therapy designed to reduce criminal justice and violence involvement among Black and Latinx youth in Chicago. Over 24 months, youth offered the program experienced an 18 percent reduction in the probability of any arrest and a 23 percent reduction in the probability of a violent-crime arrest. These statistically significant impacts, with smaller magnitudes, continue to persist up to 3 years post randomization. To better understand the behavior change we observe given an arrest is a proxy for criminal behavior, we create a supervised machine learning algorithm from arrest narratives that determines if an arrest was initiated more or less at the discretion of police. We find that the program's impacts are concentrated in arrests where officers have less discretion in initiating contact, while having little impact on more discretionary contact arrests (e.g. a young person exhibiting "suspicious" behavior). This analysis suggests the effects of the program are being driven by a reduction in youth offending behavior rather than by avoiding police contact.

VERSION: October 2024

Suggested citation: Abdul-Razzak, Nour, and Kelly Hallberg. (2024). Unpacking the Impacts of a Youth Behavioral Health Intervention: Experimental Evidence from Chicago. (EdWorkingPaper: 24-1053). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/b9dh-vq42

Unpacking the Impacts of a Youth Behavioral Health Intervention: Experimental Evidence from Chicago *

Nour Abdul-Razzak† Kelly Hallberg‡

July 8, 2024

Abstract

Racial disparities in violence exposure and criminal justice contact are a subject of growing policy and public concern. We conduct a large-scale, randomized controlled trial of a six-month behavioral health intervention combining intensive mentoring and group therapy designed to reduce criminal justice and violence involvement among Black and Latinx youth in Chicago. Over 24 months, youth offered the program experienced an 18 percent reduction in the probability of any arrest and a 23 percent reduction in the probability of a violent-crime arrest. These statistically significant impacts, with smaller magnitudes, continue to persist up to 3 years post randomization. To better understand the behavior change we observe given an arrest is a proxy for criminal behavior, we create a supervised machine learning algorithm from arrest narratives that determines if an arrest was initiated more or less at the discretion of police. We find that the program's impacts are concentrated in arrests where officers have less discretion in initiating contact, while having little impact on more discretionary contact arrests (e.g. a young person exhibiting "suspicious" behavior). This analysis suggests the effects of the program are being driven by a reduction in youth offending behavior rather than by avoiding police contact. *JEL codes:* C53, C93, I12, K40, K42

*This is a University of Chicago Crime Lab research project, with generous support by the Sports Alliance, CDC, and AbbVie. We'd like to thank Brightpoint (formerly Children's Home and Aid) and Youth Advocate Programs in making the intervention possible. We'd like to thank Cristobal Pinto, Brandon Domash, Lydia Jessup, Christina Leon, Max Lubell, Heather Bland, Lucia Delgado Sanchez, Emily He, Luke Karner, Chelsea Hanlock, Kyle Pinder, and Mariah Van Ermen for excellent research assistance and program management assistance. Special thanks to Jens Ludwig, Aurelie Ouss, Chris Blattman, Leonardo Bursztyn, Max Kapustin, Sarah Heller and seminar participants at the Transatlantic Workshop on the Economics of Crime, the Virtual Crime Economics (ViCE) seminar, Texas Economics of Crime Workshop, and the America Latina Crime and Policy Network Conference (AL CAPONE) for helpful comments and feedback. Nour Abdul-Razzak gratefully acknowledges support from the National Science Foundation (NSF) Graduate Research Fellowship Program. Research reported in this publication was additionally supported by the Centers for Disease Control (CDC) National Center for Injury Prevention and Control under award number 1-R-01-CE002971 and of the National Institutes of Health under award number 5P01HD076816. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC, the National Institutes of Health, NSF, other funders, data providers, or implementing agencies.

†University of Chicago Inclusive Economy Lab and Harris School of Public Policy. Email: abdulrazzak@uchicago.edu, Address: 111 W Washington St, Suite 1023, Chicago, IL 60602.

‡University of Chicago Inclusive Economy Lab and Harris School of Public Policy. Email: hallberk@uchicago.edu

1 Introduction

In the last decade, there has been significant progress in identifying programs that reduce violence as well as criminal justice contact in the United States.[1] Many of these programs seek to address the long-standing inequities in access to mental health resources and build on behavioral science principles, such as cognitive behavioral therapy (CBT) (Heller et al., 2017; Landenberger and Lipsey, 2005; Bhatt et al., 2023; Barnes et al., 2017). CBT-based interventions have shown to be effective in reducing crime and violence outside the United States as well (Blattman et al., 2017; Dinarte-Diaz and Egana-delSol, 2024; Blattman et al., 2023; Arbour, 2021). These types of interventions typically help individuals by slowing down reactions and learning alternative responses to stressful or challenging situations. Most of the successful behavioral health-informed youth interventions have been implemented in institutional settings (e.g. schools and detention centers). However, the same factors that cause youth to be at elevated risk for violence or criminal justice engagement make them less likely to engage key social institutions such as schools that provide access to potentially helpful programming. Youth may also have unmet needs or face barriers such as housing or food insecurity that prevent them from fully engaging in beneficial programming whether in school or not. What is missing is a way to involve the large number of young people who are beginning to disconnect from school or who are no longer in school, without having to wait until they become deeply engaged in the criminal justice system and are detained to reach them.

In this paper, we test a new behavioral health intervention in Chicago (Choose to Change, henceforth C2C) that seeks to combine trauma-informed group CBT with intensive mentorship as a mechanism to engage youth who are at elevated risk of criminal justice involvement. C2C's six-month program connects an underserved population of youth, youth facing intersecting challenges such as prior criminal justice contact and school truancy, with a comprehensive set of supports that seek to address each young person's specific needs while also building social-emotional skills. In 2015, we designed and implemented a large-scale randomized controlled trial (RCT) of the C2C program to determine it's impact, with 2,074 youth randomly assigned over four years to a treatment group that were offered C2C services or a control group that was not. This paper presents the results of the RCT using administrative arrest data, following study participants for up to three years post-randomization.

[1] A growing literature in the social sciences quantifies the high social costs of violence exposure and frequent justice system contact among youth, with consequences documented in educational outcomes, mental health, employment, and political participation (Ang, 2021; Ang and Tebes, 2020; Owens, 2017; Hjalmarsson, 2008; Flannery et al., 2004; Cloitre et al., 2009; Aizer and Doyle, 2015; Western, 2006; Legewie and Fagan, 2019; Desmond et al., 2016; Lerman and Weaver, 2014). Such costs are borne largely by youth of color (Gelman et al., 2007; Weaver and Geller, 2019; Margolin and Gordis, 2000; Ford et al., 2008; Ang, 2021).

Second, our paper also seeks to improve on the use of arrests as a proxy for criminal behavior. A critique that is made of these programs is that they are not reducing real criminal behavior, but simply helping people avoid police contact and, in turn, arrests. Given the social harms created from any engagement in the criminal justice system, this critique may be less relevant for policy makers seeking to reduce overall contact with the criminal justice system. However, understanding this distinction is useful as we seek to understand not only *if*, but *how* youth behavior is changing as a result of the program. It is widely acknowledged that an arrest is the result of both civilian and police decisions. Literature across various fields has underscored the substantial degree of discretion that police officers possess in how they enforce the law in the United States, especially with respect to proactive policing strategies that focus on surveillance and prevention (Goldstein, 1963; Linn, 2009; Nickels, 2007; Wu and Lum, 2017; Cho et al., 2021).[2] From an evaluation perspective, discretion reduces the strength of arrest data as a signal of youth offending behavior. C2C program effects (or lack thereof) may be driven by changes in civilian crime/offending behavior, civilian responses to police interactions (including youth learning to interact with police more constructively once stopped), or changes in behavior that alter the likelihood of surveillance or a police stop. To address this, we utilize rich data from arrest narratives and machine learning tools to create a new category of arrests that distinguishes between more vs. less discretionary police contact. Our paper first presents the results of the RCT on criminal justice outcomes using administrative arrest data. Second, we introduce our new category of arrests that incorporates police discretion in the context of the RCT and, in doing so, better understand how behavioral health interventions can change youth behavior.

Study youth for the RCT were drawn from neighborhoods in Chicago experiencing the highest levels of violence and youth involvement in the criminal justice system. Of the 1,052 youth offered the program, 62% took up services, a take-up rate higher than many in-school programs in Chicago. Looking at our main criminal justice outcomes, we find that C2C has a meaningful and significant impact on the extensive margin of arrests (whether or not youth have any arrests) and on arrests for violent offenses. We find that being offered C2C substantially reduces the probability of any arrest by 6.3 percentage points (18% of the control mean) during the 24 months after randomization (p-value < 0.01). Treatment-on-the-treated estimates show a 10.3 percentage point (31%) reduction in the probability of any arrest compared to the control complier mean. When we break this down by the charges associated with each arrest: violent, drug, property or "other" categories, we find a large

[2]Importantly, this discretion interacts with the racial bias that has been documented in policing (Goncalves and Mello, 2021; Hoekstra and Sloan, 2020; Antonovics and Knight, 2009; Gelman et al., 2007; Goel et al., 2016).

reduction in arrests for violent crime offenses but no other category of arrests. In the 24 months after randomization, we find that being offered C2C reduces the probability of any arrest for a violent offense by 3.7 percentage points (23% of the control mean, with p-value $<$ 0.01). Along the intensive margin at 24 months, we see that the program reduces the number of arrests for violent offenses (18% reduction of control mean, with p-value $<$ 0.10). The reduction in overall number of arrests is not statistically significant at conventional levels. The large and significant extensive margin impacts continue to persist up to 36 months post randomization. At 36 months post randomization, C2C continues to reduce the probability of any arrest by 5.2 percentage points (13% of the control mean, with p-value $<$ 0.01), and the probability of any arrest for a violent offense by 3.1 percentage points (16% of the control mean, with p-value $<$ 0.05).

These findings are very encouraging and provide policymakers seeking to reduce arrest contact and violence involvement among youth useful evidence on the value of mentorship to engage a population of youth more disconnected from school with trauma-informed therapy. The longer-term impacts of this program, with 78% of youth within the study outcome window becoming adults, are especially promising as many youth programs see fade-out shortly after program end and few studies have been able to track youth into adulthood (Heckman and Kautz, 2013; Heller, 2014; Davis and Heller, 2020).

In an attempt to better understand how this program is changing youth behavior, we dive deeper into the use of arrest data as a proxy for criminal behavior by seeking to disentangle high discretion arrests more related to police surveillance from low discretion arrests more related to actual criminal behavior. We analyze text from arrest narratives and create a supervised machine learning algorithm that seeks to identify the *context or the reason for initiation* of the police-youth interaction and determine if it was initiated for reasons that officers have more discretion over. Specifically, we designate a new category of arrest from text analysis of the full arrest record that seeks to capture one aspect of policing strategy: *high discretionary initiation*, which constitute arrests that result from contact that was initiated by police for reasons more subjective/discretionary in nature or that are part of existing proactive policing strategy, such as hot spot policing or stop-and-frisk.[3] Examples of high discretionary initiation include a police officer stopping a youth for vaguely defined suspicious behavior or being a young person present in a designated "hot-spot".[4] We find that our model can capture the discretionary variation that is present *within* specific charges or types of crimes, such as drug crimes, disorderly conduct, traffic violations and "other"

[3]Any arrest not classified as high discretionary is classified as low discretionary.

[4]Proactive policing strategies often involve preventing crime and disorder before they transpire, rather than reacting to crimes once they have occurred (National Academies of Sciences et al., 2018).

arrests, highlighting the value of moving beyond just a charge level analysis in program evaluations.[5] Ex-ante, it is not clear if C2C would impact more discretionary contact by police. While youth would seem to have less control over more discretionary police contact given the proactive policing strategies frequently used in Chicago, program youth may learn to avoid scenarios, people, or places where officers are more likely to enact discretionary enforcement. Employing this categorization of arrests, we find that the main treatment effect of C2C on arrests is driven by a reduction in the probability of *low discretionary* arrests (arrests that have little to no discretion in police initiation and often come from 911 calls or citizen complaints). Specifically, we find that being offered C2C reduces the probability of any low discretionary arrest by 18%, or 5.5 percentage points, in the 24 months after randomization (p-value < 0.01) with no detectable impact on more discretionary arrests in that time frame. A similar pattern emerges when we examine police stops data, where we find C2C does not change the number of stops or likelihood of a police stop.

Taken as a whole, these results suggest that C2C leads to a sustained change in youth behavior. We see significant reductions in the likelihood of having any arrest and any arrest for a violent-offense that persists 2.5 years beyond the end of programming. We see this behavior change in violent engagement as well as other behavior related to interpersonal conflict (e.g. disorderly conduct) that comprise our newly defined low discretionary arrest category. Importantly, violent-crime arrests make up less than a third of all low discretionary arrests, highlighting the behavior change we see among C2C youth goes beyond a reduction in violent offenses. The discretionary classification of arrests enables us to confirm that these are true changes in civilian behavior and not driven by youth responses to avoid police or detection. It also allows us to look beyond a handful of charges that are known to have little discretion (such as assault or battery) and help us understand what is driving our main finding on reducing the likelihood of being arrested. Robustness checks in the paper highlight the value of our new approach compared to existing methods (such as using index crime categorizations) of accounting for discretion in policing.

Our paper makes two main contributions to the literature. First, our paper builds on the limited but growing cognitive behavioral therapy (CBT) and behavioral health intervention literature that has demonstrated these approaches to be effective in reducing violence and arrests (Heller et al., 2017; Landenberger and Lipsey, 2005; Blattman et al., 2017; Bhatt et al., 2023). Our findings highlight the therapy + wraparound/mentoring approach can be effective in engaging a harder to reach population of youth, those who are disengaging or have

[5]For instance, our model categorizes only 22% of disorderly conduct arrests as high discretionary, underscoring the vast majority of these arrests come from interpersonal conflict or incidents that are flagged to police through 911 calls.

disengaged from school. While our study design does not enable us to disentangle the value-add of the two primary program components (intensive mentorship and CBT), qualitative data collected from C2C youth and program staff suggest that the intensive mentorship is critical for engaging harder to reach young people in CBT, and helping youth *practice* the CBT skills in reality (i.e. learning by doing). In fact, we demonstrate that the C2C program was able to engage a substantial number of youth who were not attending school consistently. Previously studied behavioral health interventions focused on youth disengaging from school have involved very small samples limiting generalizability or have found null effects (Chandler et al., 2011; Dynarski et al., 1998; Borduin et al., 1995; Larson and Rumberger, 1995). Given the relationship between disengaging in school and involvement in the criminal justice system, it is critically important from a policy perspective to have effective strategies to support this population that persist into adulthood(Herrera et al., 2013; Cunha et al., 2006, 2010).

Second, we introduce a new approach to measuring offender behaviors that builds on two strands of research literature–program evaluation of crime prevention programs and studies of police behavior–by directly operationalizing a way to distinguish between low and high discretionary arrests. Program evaluations often rely on administrative arrest records as the best proxy for offending (Doleac, 2020). And yet, these proxies have limitations given the substantial legal latitude police officers have in their work, especially with respect to lower-level offense arrests (Natapoff, 2017; Weisburst, 2022; Brown et al., 2009; Allen, 2005; Smith and Visher, 1981; Mendias and Kehoe, 2006; Rosenfeld et al., 2012; Stroshine et al., 2008). Past research using administrative arrest data has attempted to isolate civilian offending behavior from police discretion by focusing exclusively on violent offenses, which ostensibly involve less police discretion. However, while violent offense arrests are more socially costly, they also constitute a small portion of the total contact civilians have with police.[6]. Another approach researchers have used when seeking to understand all criminal behavior and still account for discretion, is to classify large categories of charges, such as non-index crimes, as discretionary (Ba et al., 2021; Rivera, 2022; Lum and Nagin, 2017). Our paper provides a more data-informed approach to identifying high-discretion arrests and specifically captures the variation in discretion that is present *within charges*, highlighting the value of moving beyond a charge level analysis. Other researchers have attempted to avoid the challenges of relying on administrative arrest data altogether by using surveys to understand criminal behavior. However this can be costly to implement and may suffer from self-report bias (Weaver et al., 2019; Blattman et al., 2016; Kling et al., 2005, 2007).

[6]In our study, violent arrests only comprise 22 percent of all outcome arrests. Nationally, fewer than 20 percent of all arrests are for serious violent or property offenses (FBI UCR 2018)

Lastly, there is a robust literature measuring police discretion by looking at officer-level differences in the propensity to engage in some kind of enforcement (vehicle searches, stops, arrests, etc) (Ba et al., 2021; Weisburst, 2022; Goncalves and Mello, 2021; Gonçalves and Mello, 2023; Feigenberg and Miller, 2022). We complement this officer-level literature by seeking to measure police discretion *for each individual arrest* using administrative data. To the best of our knowledge, this paper is the first to engage in this type of discretion measurement exercise in the context of a civilian-focused RCT. Importantly, this approach helps us understand how such programs improve the welfare of its beneficiaries and change behavior.

The paper is organized as follows: first we discuss the details of the Choose to Change intervention and the theory behind how the intervention was designed to support young people. Next, we discuss the study population and the data used in the project. We then describe the methods behind the RCT and the main RCT findings. Lastly, we dive into the discretionary mechanism analysis, and a discussion of findings.

2 Choose to Change: Program Model

In response to a 2015 Design Competition, an initiative launched by the University of Chicago Crime Lab and Education Lab to crowdsource youth violence prevention interventions from across the city of Chicago, two local Chicago nonprofits, Brightpoint (formerly Children's Home & Aid) and Youth Advocate Programs (YAP), Inc. created Choose to Change (C2C): Your Mind, Your Game.[7] C2C offers youth a program combining individualized advocate-wraparound services, including intensive mentoring, with group-based, trauma-informed therapy. This combination supports youth by first helping them understand how past traumatic experiences and chronic stress can impact their thinking and behavior and how that, in turn, affects their emotional responses to situations they encounter in daily life. Over the course of the six-month program, advocates from YAP meet with youth individually for at least eight hours a week, building strong interpersonal bonds and offering wraparound and mentoring services focused on highlighting each young person's strengths and addressing their specific needs. In addition, therapists from Brightpoint lead trauma-informed CBT sessions called SPARCS (Structured Psychotherapy for Adolescents Responding to Chronic Stress), which helps youth process prior trauma and develop a new set of decision-making tools that seek to reduce maladaptive responses and behaviors. Program staff identify four core components that they credit with the program's success: relentless, strength-base engagement;

[7]The 2015 Design Competition was held in partnership with GET in Chicago, a local philanthropic organization and the MacArthur Foundation.

applied learning; developing decisions making skills; and building strong relationship. We briefly describe each of these program components below and summarize the main program elements in Tables A.I and A.II in the Appendix.

2.1 Relentless, Strength-Based Engagement

Youth are referred from either a community-based partner, public agency, or their middle or high school based on the referral partners' assessment of their risk for engagement in violence or the criminal justice system. C2C advocates then connect with the youth and their family members to introduce the program and obtain consent to participate. Advocates are persistent in their recruitment and engagement of young people, and try multiple tactics to locate youth including using their connections in the community. This relentless engagement pays off with 62% of youth agreeing to participate.

Services formally begin with the advocate convening a youth family meeting that includes the youth, their parents, other family members, and any other individuals who might be able to support the youth. Through this meeting, the advocate develops a picture of the youth's life and how they can advocate on the youth's behalf to help them reach the best version of themselves. The advocate, with the cooperation and input of the family and youth, then draws up a service plan (including action steps and goals) based on the improvement areas and strengths identified. These goals could include a young person's desire to get employment, address legal challenges, help secure stable housing or educational aspirations.[8] Because the wraparound services are tailored to the individual, what services are provided is dependent on the needs of the youth and their family.

This strengths-based approach continues throughout a young person's engagement in the program. C2C employs a "No Reject, No Eject" policy that dictates that youth are not denied or discontinued from services in response to their life circumstances, new challenges, or for non-compliance with program policies. Rather, C2C staff work to meet youth where they are and adapt services to meet their specific needs and situations. Throughout the intervention, the advocate engages in one-on-one meetings with the youth, family meetings, and weekly recreational groups. The group activities tend to be fun experiences, such as playing basketball at a local gym, going to see movies, or going out to eat. These interactions also serve as an opportunity for advocates to get to know their clients and build trust. Advocates provide around-the clock support to the young people they serve, communicating frequently via text message or phone calls between formal outings.

[8]The model is rooted in the belief that all youth, adults, and families have strengths that can and should be developed. The principle of strength-based services encourages teams to create goals that reflect building family and youth assets, capacities and resilience, rather than focusing on deficits and problems.

2.2 Developing Decision Making Skills

The C2C program utilizes the Structured Psychotherapy for Adolescents Responding to Chronic Stress (SPARCS) curriculum to support youth in developing the skills needed to disrupt negative thoughts, resolve conflicts, and build self-efficacy (De Rosa et al., 2004). Typically, groups of eight to 10 youth will attend 12-16 sessions, 45-60 minutes each. Once sessions start, they are held once a week and typically scheduled in school (to incentivize attendance) or at Brightpoint's facilities. SPARCS targets six domains of functioning: regulating emotions and behavior, attention and awareness, self-concept, relationships, physical complaints, and hopefulness and sense of purpose in life. In SPARCS, participants learn how to recognize stress in the body, utilize coping skills in the moment, identify alternative ways to respond to achieve their goals, and develop communication skills.

SPARCS is trauma-informed and aims to help youth understand and change the way stress or past traumatic experiences can influence their decision-making. Therapists stress how these "emotional leftovers" from bad experiences can make youth more willing to engage in short-term coping strategies that can lead to worse outcomes or escalating conflict. SPARCS helps youth unlearn destructive behaviors, with an important emphasis on identifying triggers for dysfunctional behaviors and the circumstances that are maintaining these behaviors (Van Dijk, 2013). Through exercises and conversation, youth learn to better regulate their emotions, engage helpful coping strategies, and build problem-solving and communication skills. SPARCS combines elements from traditional CBT and Dialectical Behavioral Therapy (DBT), a form of CBT that incorporates more mindfulness and acceptance techniques.[9]

2.3 Applied Learning

The SPARCS sessions are consistently presented to youth in a non-clinical framework. For example, group-based therapy sessions are led by Masters-level clinicians, but they are referred to as coaches rather than therapists to ensure that they are approachable to youth who may have misgivings about participating in therapy. Importantly, the program model creates many synergies that help reinforce skills for the youth. During the program, advocates are required to attend the SPARCS sessions with the youth. Advocates then reinforce

[9]The curriculum incorporates core elements of evidence-based treatments for youth experiencing trauma, including psychoeducation (developing an understanding of how people react to trauma), relaxation and emotion regulation skills, mindfulness, and cognitive skill building (Santiago et al., 2018). An important element of SPARCS, and DBT more generally, is removing judgement of current maladaptive behaviors (Van Dijk, 2013). The manual reminds the therapists that even maladaptive coping skills are coping skills, and youth are doing the best they can with the skills they have.

the SPARCS skills in their individual and group interactions with youth out in the community. This allows for "learning by doing" and creates opportunities for youth to develop new habits that will benefit them after the program ends. Furthermore, since the advocates interact regularly with program participants in various settings, they can highlight key moments when SPARCS techniques can be useful. Youth often call or text their advocate if a stressful or safety issue arises. Advocates can use these opportunities to remind the youth of helpful coping strategies. This is strengthened as the therapists and advocates try to maintain a close relationship to track the progress of the youth. Advocates also provide a warm handoff to the therapy component of the program, as the group CBT sessions do not start until a month after the program has started to ensure some trust has developed between the young person and the advocate.

2.4 Building Strong Relationships

The final pillar of the program is the strong relationships advocates and coaches build with the young people they serve. C2C staff bring compassion and lived experience to building these relationships. Advocates and youth share similar backgrounds often coming from the same neighborhoods, which helps youth develop functional and trusting relationships that are key to helping them continue to engage in the program and the CBT sessions. Advocates serve between five and 15 youth at a time to support relationship building. One key aspect of the model is to build the social capital for the youth and create an engaged and sustainable family team working in partnership with the advocate and the therapist.

The bundled treatment cost about $5000-7600 per youth at the time the program was implemented, with the cost increasing over time in response to increasing wages. Because we study the complete bundle of program services in their entirety, we are only able to identify the effect of the program as a whole and cannot disentangle the relative contribution of the CBT or mentoring components or and synergies generated by implementing these program components together.

3 Eligibility and Study Population

Since 2015, the research team has worked closely with the program providers to reach out to referral agencies able to identify eligible youth who might benefit from C2C. The study population consists of $2,074$ young people drawn from neighborhoods on the south and west sides of Chicago that have been exposed to challenging conditions including but not limited to segregation, poverty, disinvestment, and state and interpersonal violence. Since

the C2C advocates frequently travel to the youth's place of residence and school, geographic boundaries were drawn at the start of the program in order to ensure advocates were not driving excessive distances to reach youth. Figure A.I shows a map of where study youth reside based on their address of residence at the time of referral.

The neighborhoods from which the study youth are drawn experience some of the highest rates of violence and contact with police in the city. Figure A.II shows a map of the 5-year average number of shootings by community area (2015 to 2019) and Figure A.II shows the number of complaints against CPD in the last three decades (using data from the Invisible Institute). Both maps depict high concentration in the south and westside neighborhoods where our study youth reside.

For all cohorts, in the months preceding engagement with youth, C2C staff conducted outreach to community service providers and schools in the neighborhoods they served to identify and recruit youth who would meet the target population for this program. Specifically, referral sources were asked to identify youth between the ages of 13 and 18[10] who were: actively gang-affiliated or at risk or gang engagement; on juvenile probation; previously found guilty of weapons offenses;seriously disruptive in school through chronic truancy, serious misconduct and/or frequent suspensions; and/or direct victim of or witness to traumatic violence. Our study team received referrals from Chicago Public Schools (CPS) - including neighborhood and charter high schools (56%), alternative schools (15%), the Student Outreach and Re-engagement Centers (20%) focused on re-engaging chronically truant and out-of-school youth and the Office of Safety and Security (6%) - as well as Cook County Juvenile Probation (3%). Administrators would identify the youth in need of services based on their internal data and the referral criteria provided to them. Program providers preferred this approach to a data driven approach given the knowledge and relationships school administrators had with youth.

Randomization and enrollment were done on a rolling cohort basis over four years, but all youth were ensured a minimum length of programming. For the first four cohorts, youth were enrolled for five months of services. This expanded to six months in cohort 5 as more funding became available. Randomization and subsequent enrollment continued until a cohort had reached capacity (typically 100 youth, although this varied cohort to cohort depending on funding and decreased over time with later years serving about 50 to 60 youth per cohort). We use the randomization date as the start of the post-randomization outcome period, but recognize that treatment youth may have actually enrolled in the subsequent weeks or even

[10]Although the program targeted 13-18 year-olds with respect to the program letters and advertisement sent out to the referral agencies, occasionally youth outside of this age range were served if age was not available prior to randomization.

months. The RCT includes eight cohorts of youth, beginning randomization in November of 2015 and wrapping up randomization by December of 2019. The majority of the youth in the last cohort of programming (cohort 8) were in services during the early days of the COVID-19 pandemic.[11] Across all eight cohorts, 1,052 youth were randomized to treatment and 1,022 youth were randomized to control. Randomization was stratified by referral source.

Table I shows that the program succeeded in identifying youth in the target population. Referred youth were, on average, almost 16 years old and in the first few years of high school. About 95% of the youth are Black and almost all youth qualify for free or reduced lunch. A significant percentage had been previously arrested (35%), and 20% had a gap in school enrollment at some point during the prior year. Randomization was successful in balancing the treatment and control groups across almost all baseline characteristics (see Table I). An F-test for joint significance shows we cannot reject the null hypothesis that treatment and control groups are equal. We recognize one category of arrest is marginally significant in the difference between treatment and control groups within both the intensive margin of arrests and extensive margin of arrests. Given the large number of baseline characteristics we look at, by chance it's very possible to find one or two characteristics that are not perfectly balanced. Given the F-test, we are not concerned with this. We also control for all of these baseline characteristics in our regressions.

4 Data and outcomes measured

4.1 Criminal justice outcomes

To understand how C2C impacts contact with the justice system, we matched our study youth to Chicago Police Department (CPD) arrest data using a probabilistic match based on first name, last name, and date of birth.[12]. For our main outcomes, we show both the number of arrests after randomization, and whether or not a youth has any arrest after randomization. Given that the number of arrests variable is skewed with a mass at zero, an analysis of the average change might miss important changes in whether youth had

[11] Given that most of the randomization and recruitment happened prior to COVID, we do not see significant impacts on the take-up rate during this cohort. However, programming obviously changed dramatically with most services moving online. Necessities such as groceries were still delivered to families in person, but mentoring and therapy services were mainly conducted virtually, with some in person outdoor services occurring.

[12] The Appendix discusses details of the matching procedure. Our project through the University of Chicago Crime Lab has a master data sharing agreement with the Chicago Police Department and Chicago Public Schools. The CPD data was provided by and belongs to the CPD. Any further use of the data must be approved by CPD. Points of view or opinions contained within this document are those of the authors and do not necessarily represent the official position or policies of the Chicago Police Department

any engagement in the criminal justice system through being arrested. We believe both margins are critical to examine as any arrest can have negative consequences for youth and society (Aizer and Doyle, 2015; Western, 2006; Legewie and Fagan, 2019; Desmond et al., 2016; Lerman and Weaver, 2014; Gelman et al., 2007; Weaver and Geller, 2019). For the main RCT results, we follow the existing literature and break down number of arrests by charge type (violent, property, drug, or other), with the "other" arrest category consisting of arrests for violations such as trespassing, disorderly conduct, weapons violations, vandalism, warrants, etc.[13] Violent arrests comprise about 22 percent of the outcome study arrest sample, property arrests make up about 24 percent, drug arrests make up 7 percent and other arrests make up almost 45 percent of the arrests in the sample. Treatment and control group differences are assessed every six months post randomization. Our pre-registered primary outcome (with AEA's RCT registry) is number of arrests for the full study up 24 months post randomization, looking at outcomes at 6 month intervals.[14] Our other pre-registered primary outcome is school engagement, however due to data delays and the need allow for enough time to pass to allows the full study sample the opportunity to graduate from high school this analysis cannot be complete at this time. A future paper will report on the educational outcomes as well as a cost-benefit analysis for the program.[15]. Our pre-registered secondary outcome was looking at arrests broken up by the different charges associated with each arrest. At present, enough time has elapsed to track the entire sample of 2,074 youth for 36 months post randomization, and 1,662 youth (80% of the sample) for 48 months post-randomization. We will focus on the time periods for which we have outcome data for the full study sample. Examining these time intervals provides some insight into whether the program succeeded mainly due to incapacitation–by keeping the youth busy during the first six months post randomization while the youth are in programming–or if the effects persist once the program has ended.

One important note about our outcome data is that we rely on administrative CPD data. Therefore we do assume that if a youth is not matched to CPD they essentially have "zero" arrests or stops. However, if youth move out of Chicago, this could represent an undercount. This would be particularly concerning if this measurement attrition was differential across treatment condition. To assess whether study youth moved out of Chicago at differential rates, we use Chicago Public Schools (CPS) transfer data given that we have baseline CPS

[13]Violent arrests comprise of homicides, sexual assault, robbery, aggravated assault, aggravated battery, simple assault and simple battery.

[14]`https://www.socialscienceregistry.org/trials/933`.

[15]Many younger youth were enrolled in the last cohort in 2019, and therefore we need to wait and observe a potential graduation (allowing for an extra year) for them in 2024. Employment effects were the other pre-specified outcome that will not be possible due to data limitations.

data for 99.8% of all study youth. CPS records all youth who transfer, and in particular note if they have transferred outside of Chicago in enrollment data. Using this indicator, we find no evidence of a differential rate of departure for those assigned to treatment compared to the control group and the rates of moving out at all are small even up to 18 months post randomization (see table A.III in the Appendix).[16]

4.2 C2C Program Take-up and Dosage

Table II provides a summary of program take-up and dosage. Of those assigned to treatment, 62% enrolled in the program.[17] We define take-up as participating in at least one trauma-informed CBT session (SPARCS) or receiving at least five hours of mentoring/advocacy services. To put this take-up rate in context, we compare C2C to the Becoming a Man (BAM) program that was evaluated by (Heller et al., 2017) in Chicago and saw a take-up rate of 50% (defined as attending at least one CBT session). It is noteworthy that C2C was able to achieve a higher take-up rate even though BAM is a school-based program that operates during the traditional school day and serves young people experiencing fewer risk factors for criminal justice system engagement. [18]

In fact, the C2C program was able to engage young people who were experiencing risk factors for criminal justice system engagement at relatively high rates. Of the 206 youth assigned to treatment who had a gap in school enrollment prior to randomization, 45% took up the program. Likewise, 51% of youth who had a baseline arrest and 46% who had a baseline violent arrest enrolled in the program. To underscore the risk characteristics of youth in C2C, some youth in the C2C study would not have been eligible for in-school programs like BAM because of their limited school engagement. Specifically, we calculated that between 221 and 377 of the C2C treatment group (21% - 36% of the treatment group) would *not* have been eligible to participate in the BAM program based on their previous limited school engagement. [19] Of these young people, C2C still saw take-up rates between

[16] 18 months is the longest time outcome period we have for all study youth due to data delays. Specifically, we find that in the 18 months after randomization, study youth spend an average of 32 days not in Chicago. Given that youth can move out of Chicago temporarily and then move back into CPS, we believe looking at the total number of total days they spend not enrolled in Chicago is a more accurate measure of this censoring issue.

[17] There are 4 youth assigned to the control group who engaged in C2C services, for a total of 657 youth in the program during this time period.

[18] Directly comparing the baseline characteristics of the BAM study (2009-2013) participants to C2C study participants may not be useful given the overall decline in juvenile arrests over time in Chicago. Regardless, we see that C2C youth are similar to BAM's study 1 youth (2009) on many baseline characteristics, and C2C youth are *higher* risk on many dimensions (including arrests) when compared to BAM study 2 (2013).

[19] Because the BAM program is administered during the school day, students had to be present to benefit. For that reason those who missed more than 60% of days, received a grade of "F" in at least 75% of their

44% and 47%, highlighting C2C's ability to not only offer these youth helpful programming, but also engage them.

Youth who enrolled in the C2C program on average spent 186 hours receiving services, indicating very intensive engagement. This included an average of nine hours in SPARCS sessions and 177 with their mentor (either one-on-one or in a group setting). For comparison to a school-based program, BAM participants averaged 13 hours of service in the first year of the study and 17 in the second year of the study. Even those in the 25th percentile of engagement for C2C attended four SPARCS sessions and received 126 hours of mentoring support. The distribution of dosage in SPARCS does indicate that youth with more risk factors on average engage less in SPARCS, reinforcing the need for a program model like C2C. Like overall take up, we saw a similar level intensity of engagement among youth with multiple risk factors for criminal justice system engagement. Almost all of the young people who were served by the program received both mentoring support and attended SPARCS sessions (82% of those who enrolled in the program engaged in both program components).

Alternatively, we could have defined participation as an individual who receives at least one CBT/SPARCS session *and* five hours of mentoring. This would have resulted in a take-up rate of 51%. Given the program model and how mentors are also reinforcing CBT skills (in the 7 hours a week on average they spend with youth), we prefer to use the more inclusive measure of take-up. Over time, as the program worked out implementation issues, more youth received SPARCS sessions.

While our study design does not enable us to disentangle the value-add of the two primary program components (intensive mentorship and SPARCS), qualitative data collected from C2C youth and program staff suggest that the intensive mentorship is critical for engaging harder to reach young people in CBT. Drawing on interviews with front-line staff and focus groups with youth participants, we found that many of the young people who were assigned to C2C were initially skeptical both of the program generally and participating in the CBT sessions in particular. The persistent engagement and the social ties built with the program mentors were critical to overcoming this reticence and supporting engagement with the CBT sessions.[20]

courses, had a serious IEP, or were at least two years older than expected for the grade they were enrolled in were not eligible for the program and study. Drawing on administrative data from CPS, we produced more and less conservative estimates of the portion of our sample that would have been disqualified accounting for missing data.

[20]A working paper discussing the extensive qualitative work will be released soon with Max Lubell, Sociology PhD student at the University of Texas at Austin.

5 RCT Analysis Approach

To estimate the intent to treat (ITT) effects, we run the following regression:

$$Y_{ijt} = \alpha + \beta^{ITT} Z_{ij} + \gamma X_i + \phi_j + \epsilon_{ijt} \tag{1}$$

Where Y_{ijt} represents the outcomes of interest during the post-randomization period t for individual i, in block j, Z_{ij} represents the random assignment indicator for each youth within each block (j). X_i are the baseline covariates included to improve precision by accounting for residual variation in the outcome of interest, and, ϕ_j are the set of dummy variables indicating the observation's randomization block (by referral source and cohort). β, the ITT captures the impact of being offered the C2C program, and may be more relevant for policy purposes. However, given that not everyone who is offered the program participates, we present the effects of participating in the program as well.[21]

To estimate the effect of participating in the program, we estimate the effect of the treatment on the treated (TOT) using a two-stage least squares instrumental variables approach that treats random assignment as an instrument for participation, as follows:[22]:

$$P_{ij} = \alpha_1 + \beta_1 Z_{ij} + \gamma_1 X_i + \phi_j + \epsilon_{ij1} \tag{2}$$

$$Y_{ijt} = \alpha_2 + \beta_2 P_{ij} + \gamma_2 X_i + \phi_j + \epsilon_{ijt2} \tag{3}$$

Where P_{ij} indicates participation in the program by a youth in a block after randomization. β_2 approximates the TOT. [23] To benchmark the TOT, we calculate the control complier mean (CCM), or the outcome mean for those who would have taken up treatment had they been offered it. This is calculated by taking the mean of the outcome for those that comply with the treatment and subtracting the TOT (Katz et al., 2001).

[21]Most of the outcomes we seek to learn about involve variables that take only a limited number of values (e.g: binary variable for being arrested at all or number of arrests within a time frame). Some may argue that nonlinear models such as probit and Tobit are preferred in these cases when the outcome of interest in not continuous. All of our main results will use ordinary least squares given that OLS gives us the average causal effect without additional distributional or functional form assumptions. Likewise, we know that OLS will also always give us the minimum mean squared error linear approximation to the conditional expectation function, and IV will capture the local average treatment effect even in the cases where we have dependent variables that take limited values (Angrist and Pischke, 2008).

[22]This analysis requires the typical relevance and exogeneity assumptions of instrumental variables. In order for the random assignment variable to be a valid instrument, it must be correlated with program participation and uncorrelated with observables. It must shift participation in a uniform direction across people (the monotonicity assumption).

[23]Given the minor control crossover – 4 youth – this is technically the local average treatment effect (LATE) but given the very low rate of crossover this should be very close to the TOT.

6 RCT Results

Figures I shows the overall arrest Intent-to-Treat (ITT) estimates for both the extensive and intensive margin over time. We find that being offered C2C substantially reduces the likelihood a young person has any arrest by 6.3 percentage points or 18 percent of the control mean in the 24 months after randomization (p-value < 0.01). TOT estimates (highlighted in Table IV) show a 10 percentage point decrease or a 31 percent reduction in the probability of arrest compared to the control complier mean over the same period. The ITT treatment effect is significant at the 1 percent level as soon as 12 months post randomization, peaks in terms of largest magnitude at 24 months post randomization, but maintains significant, large effects even at 36 months post randomization, with an impact of 5.2 percentage points, or a decrease of 13% of the control mean (see Tables III and IV). The intensive margin effect of C2C appears to be noisier and less robust. While we observe consistently negative coefficients for program effects on number of overall arrests, these impacts are not statistically significant at conventional levels and have wide confidence intervals during most outcome periods.

When we break these overall arrest impacts down by the charges associated with each arrest, we see large reductions in arrests for violent crime offenses. Figure II highlights the ITT results for violent-offense arrests over time. Twenty-four months after randomization, we find that being offered C2C significantly reduces the probability of any arrest for a violent offense by 3.7 percentage points (p-value < 0.01), or 23 percent of the control mean (16% of control youth have an arrest for a violent offense two years after randomization). TOT estimates (in Tables III and IV) show a 39 percent reduction in the probability of any violent arrest. Importantly, we see persistence in these impacts over time as youth in the study age into adulthood. We see the reduction in violent offenses immediately after the program wraps up at six months post randomization, and persist even up to 36 months post randomization. Three years post randomization, we find that being offered C2C significantly reduces the probability of any arrest for a violent offense by 3.1 percentage points or 16 percent of the control mean. Within this three outcome window, roughly 78% of the study population reaches the age of 18 and youth are 18 or older on 48% of person-days during the period. Along the intensive margin, we see similar impacts and magnitudes although slightly noisier estimates over time. Six months post randomization, we find that being offered C2C lead to a reduction of roughly two arrests for violent offenses per 100 youth (p-value <0.05).

Tables III through IV also highlight the impacts of C2C on other typical charge categories of arrests: property, drug, and other (anything that does not fall into violence, property or drug). We find that C2C does not have any consistent impact on any other type of arrest besides violent. We do see some evidence that youth who are offered a spot in the

C2C program are slightly more likely to be arrested for a drug offense while they are in the program (6 months after randomization), but this does not remain significant after correcting for multiple hypothesis tests.

Generally, all the extensive margin estimates for overall arrests and violent arrests remain statistically significant at conventional levels after correcting for multiple comparisons using the family-wise error rate (FWER) up to 24 months post randomization (Westfall and Young, 1993). FWER is defined as the chance that at least one of our outcomes in the "family" of outcomes is significant when the null hypothesis of no effect is true. We consider our family of outcomes to include all arrest types in a given time frame (6 months, 12 months, 24 months, etc) by margin (extensive or intensive).[24]

Had we stopped here, as is typical of existing program evaluations using criminal justice administrative data, we would have concluded that the program mainly operates on the extensive margin, reducing the probability of having any arrest, with consistent reductions on both intensive and extensive margins for violence. This is very encouraging as any contact with the criminal justice system can be harmful to youth, even if it's not related to actual criminal behavior on the part of the youth, as the costs imposed on the individual and system can be detrimental. Furthermore, the large reduction in the probability of a violent arrest is particularly promising given the high cost of violence for both individuals and society as a whole (Gobbo, 2023; Chalfin, 2015; Council et al., 2011).

Further disaggregating the reduction in violent offenses, we see these reductions are driven mainly by a decline in arrests for aggravated assaults and batteries.[25] We also find some suggestive evidence that C2C reduces the likelihood a young person is the victim of a violent incident. Specifically, we find that being offered C2C reduces the likelihood a young person experiences a serious violent incident by 3.6 percentage points in 36 months post-randomization, or 21% of the control mean (p-value $< .05$, see Table A.IV in Appendix).[26] We did not pre-specify this outcome given that victimizations are self-reported to CPD and involvement in C2C could lead to differential reporting rates between the treatment and control groups that could drive differences in outcomes. However, we believe these reductions are promising because violent incidences are less likely to be subject to differences in reporting

[24]FWER uses a bootstrap resampling technique that simulates data under the null hypothesis. Within each permutation, we randomly reassign the treatment indicator with replacement and estimate program impacts on all five of our main outcomes (all arrest categories in each time period). By repeating this procedure 5000 times, we create an empirical distribution of t-statistics that allows us to compare the actual set of t-statistics we find to what we would have found by chance under the null.

[25]Details provided upon request.

[26]Serious violent victimizations include part-one violent incidents which are homicide, shootings, sexual assault, robbery, aggravated assault, and aggravated battery.

to CPD. Specifically, these results, combined with the reduction in violent arrests, give a strong indication that C2C is effective at reducing actual violence involvement.

While these results are promising, we believe they may obscure a fuller picture of how the program is changing youth behavior and youth interaction with the criminal justice system more generally. Program effects (or lack thereof in certain categories) we find may be driven by changes in actual civilian behavior, civilian responses to police interactions, or changes in behavior that alter the likelihood of surveillance or a police stop. Given that youth in C2C may be changing not just their behavior but also with whom and where they hang out, it is unclear ex-ante what might be driving our overall arrest results. To better disentangle these potential mechanisms, we attempt a novel approach to measuring police discretion and examine how police discretion interacts with the effects of the C2C program in the following sections.

7 Understanding the Mechanisms Behind C2C Impact: Police Discretion

7.1 Defining Discretion using Arrests Narratives

In this section of the paper, we classify arrests by the reason for the contact that ultimately led to an arrest, and if police used discretion for the initiation of the contact. It is well established that arrests incorporate not just civilian behavior but policing behavior as well, with a long history in Criminology and other fields that documents the discretion present in policing (Goldstein, 1963; Linn, 2009; Kelling, 1999; Wu and Lum, 2017; Cho et al., 2021). Researchers have used various definitions of discretion including variation in work-related decisions (i.e. arrest, diversion or citation), variation in the use of force, and the use of extra-legal factors, such as race of suspect, in decision-making (Nickels, 2007; Skogan and Frydl, 2004; Mastrofski, 2004). When attempting to account for this discretion in arrest administrative data, some researchers have de-emphasized minor charges that are more likely to be a result of police discretion such as drug offenses or choose to focus solely on charges such as violent offenses that are less likely to involve police discretion (Lum and Nagin, 2017; Ba et al., 2021). However, these are also proxies in that charge type may not capture the discretion that comes earlier in the interaction at initiation. For example: a disorderly conduct charge can result from a call for service from a school where a fight has broken out or from a gang loitering stop where police officers are trying to clear a "high-crime" corner. Violent arrests, while socially costly, makeup only 22 percent of all our outcome arrests among the C2C study youth. In this paper, we attempt to go beyond the charge

level classification by incorporating new data–text from arrest narratives and new methods – machine learning (ML). This approach will allow us to incorporate information about the level of policing discretion in how we understand arrests as a measure of youth criminal behavior more broadly, and help us understand what is driving our main C2C finding that the program reduces the likelihood of being arrested.

It also worth highlighting there are several areas where police discretion plays a role in the police-civilian interaction, including but not limited to: the decision to monitor an area, the decision to stop someone or further investigate a situation, the decision to arrest a person (versus give a citation or issue a warning), the decision to use force, and in the number and types of charges filed. In this paper, we focus on the decision to initiate contact with a civilian.[27].

To create this new classification of arrest focused on police discretion, we first use detailed text arrest narratives to understand what led to the initiation of the arrest. Specifically, we define a *high discretionary contact arrests* as an arrest that was initiated for reasons more subjective in nature or that are part of existing proactive policing strategy such as hot spot policing or stop-and-frisk. Examples of discretionary contact include a police officer stopping a youth for vaguely defined suspicious behavior or for being in a high-crime area with recent activity.[28] Anything not considered a high discretionary contact arrest is considered a low discretionary arrest.

Almost every arrest in our study sample is accompanied by a paragraph detailing the nature of the arrest.[29] Each narrative begins with the context and details of the initiation and why the police were present in that given area (service call, patrol, etc). For example, the arrest narrative will detail if the police are responding to a call for service or a civilian complaint, or if they were on patrol and decided to stop a particular person. It then details the nature of the interaction and how the civilian reacted to the efforts of the officer. Lastly, it details the resolution of the incident and how and where the civilian was taken into custody. The arrest narratives are written from the perspective of the officer. Narratives may include intentional or unintentional inaccuracies, and we do not claim these descriptions of arrests or stops are an objective accounting of exactly what occurred in the interaction between police

[27]Future work will focus on understanding the discretion that occurs once a stop happens, and if and how C2C impacts the likelihood a stop results in an arrest (including how the interaction unfolds)

[28]Proactive policing likewise captures discretionary behavior by police officers: this strategy involves engaging in policing as a means of preventing crime and disorder before they transpire, rather than reacting to crimes once they have occurred (National Academies of Sciences et al., 2018).

[29]There are a handful of cases where the narrative is very short and does not describe how the contact was initiated. In the manual classification part of our ML work, we used the default classification of non discretionary. However, the ML model would have predicted these based on all the other non-text features included in the model, including charge information, etc.

and a young person. While the officer-youth interaction may be impacted by treatment, we believe it's reasonable to assume any potential inaccuracies in the description of why that contact was made is not correlated with treatment. As a indirect test of this assumption, we do show that narrative length or complexity is not correlated with treatment status (Table A.V).

We used a machine learning methodology to predict our object of interest (high discretionary arrests) in a model that performs well out of sample. While presumably manual text analysis is possible, it would be cost prohibitive and time consuming to manually classify all 6,364 arrests in our baseline and outcome study samples. We also think this approach can be useful in other contexts for police research or program evaluations that are seeking to classify discretionary arrests consistently, accurately, and efficiently according to arrest narratives. To employ a supervised ML model to classify the full set of arrests, we first classified a portion of arrests manually to serve as our train and test sets for the supervised machine learning model. The details of our manual classification can be found in Appendix, section A.1.4.1. The protocol for our manual classification was based on existing literature and a report from the Department of Justice that documents the circumstances and scenarios where Chicago police officers are more likely to use discretion.

Among the sample of manually classified arrests, 25.2% were classified as high discretionary.[30] Section A.1.4.1 in the Appendix highlights a few examples of arrest narratives and how they were classified manually based on our protocol. Typical examples of low discretionary arrests include officers being called to school incidents, officers explicitly looking for someone at a house with an active search warrant, officers getting waived down by a complainant, 911 calls, and officers responding to a theft at a store. Examples of high discretionary arrests include stopping a young person for a "youthful appearance" during school hours, someone looking suspicious by turning away from an officer with their hand on their waistband, or a person or a group of people spending time at a particular location in a designated "hot-spot". It is also worth highlighting that many of these high discretionary arrests occur because of the high level of police presence in many majority Black neighborhoods (Hinton and Cook, 2021). For example, gambling offenses or Chicago municipal code violations which include offenses such as riding a bike on a sidewalk are often only enforced in areas where police are present to observe such actions. In the subsequent section detailing the results of the ML model, we will discuss more details of the classification created.

[30]We suspect most of the low discretionary arrests are coming from calls for services or 911 reports. This percentage is in line with a recent paper that found on average 64 percent of arrests by the average officer are initiated by a civilian call for service (Chalfin and Goncalves, 2020).

7.2 Predicting Discretion using Machine Learning

Once our train set was created using manual classification, we then trained a machine learning model on this data. Specifically, we employed a random forest algorithm, which is a tree-based classification method that avoids overfitting by averaging predictions from many trees that have been grown from a random subset of predictors.[31] The main features or predictors of the model included text from arrest narratives (tuned text feature subset), demographic information on arrestee (race, gender, etc), geographical information (arrest beat, CPD district, etc) and information related to the charges of the arrest (top charge, number of charges, etc). Our model performs well on out-of-sample validation measures and a host of robustness checks, discussed in detail in the Appendix.[32]

Overall, our ML approach found that about a third of post-randomization arrests for our study sample (32.9%) resulted from police-youth contact that was initiated with high discretion from the officer. The last column in Table V presents the portion of arrests resulted from contact that was initiated at the discretion of the officer by FBI category of the top charge associated with that arrest and the share of outcomes that category represents. For example, all or almost all aggravated assaults, burglaries, homicides, and sexual assaults were identified as having resulted from the act of a civilian, such as a call for service, or other circumstance in which police discretion in initiation played a minimal role.[33] For these categories of infraction, charge type as typically employed may be a reasonable proxy of officer initiation. On the other hand, we are not surprised to see that categories of offenses associated with drugs, gambling, or Chicago municipal violations are overwhelmingly more discretionary. It's also important to note that the majority of arrests for weapons violations (which typically involves unlawful gun possession) are classified as high discretionary given these behaviors are typically discovered from street or traffic stops, where officers have high discretion in initiating.

However, there are several instances in which there is substantial variability *within* charge type. Digging into the narratives of these arrests show that the algorithm is picking up on real differences in behavior that would be obscured by only looking at arrest data by charge. For

[31]Tree-based classification methods essentially sort data observations into bins based on values of the predictor variables. This segments the data set into rectangular regions, and forms predictions as the average value of the outcome variable within each partition. Regression trees have become a popular nonlinear approach for text analysis as they incorporate rich interactions and nonlinear dependencies into classification in a simpler and interpretable way (Breiman et al., 1984; Breiman, 2001; Gentzkow et al., 2019).

[32]The model exhibits a true positive rate of 90% and a miss-classification rate of 16% in the test set.

[33]There are a small portion of aggravated assaults and batteries that are classified as high discretion. Diving into these arrests, we see that these are situations where a discretionary stop like a minor traffic incident or a street stop escalates into an incident where the person being arrested acts in a violent manner towards the officer or insults the officer (e.g. spitting on the officer or displaying the finger).

example, about a fifth of arrests for disorderly conduct were identified as initiated at officer discretion with the remaining arrests being flagged as low discretionary. In one example, a disorderly conduct arrest narrative that was flagged as officer initiated included a young person being flagged as being in a "hot spot" and "loitering on the corner with several other individuals in front of a building in which he does not reside, with at least one other gang member, in an area known for heavy gang or narcotic related activity." When the young person was initially approached, by the police, he voluntarily left the area, but when the police circled back, he had returned and was subsequently placed in custody and charged with disorderly conduct. By contrast, a disorderly conduct arrest that was classified as low discretionary resulted from a call of a battery in progress. In this other example, the arresting officers arrived at the scene and observed "the arrestee along with a large group fighting in the street... endangering his safety and others while disrupting the flow of traffic. The arrestee along with a large group of males ignored the arresting officers verbal command to stop and return to the sidewalk. When the arresting officers exited the vehicle to stop the verbal and physical altercation, a black object resembling a pistol fell from one of the subjects and hit the ground making a metallic sound consistent with a firearm. Said object was recovered... and found to be a gas operated replica BB gun... the subject fled from where the BB gun was recovered, running through groups of citizens on the sidewalk, recklessly and forcefully endangering the safety of the citizens and himself." The charge associated with this incident was also disorderly conduct, but police discretion in initiation seems to play a much less consequential role in the police/youth interaction. It is worth noting that disorderly conduct arrests are often used as a proxy for high discretionary arrests when researchers are trying to proxy discretion using only the charges associated with each arrest (Dube et al., 2023). Yet our ML analysis flags that in fact the majority of these types of arrests come from scenarios where officers have little discretion. One recurring theme we find present in arrests with low-discretion is that they often include some type of interpersonal engagement or interaction with other youth.

It is also worth highlighting that not all low discretionary arrests come from a call for service or 911 call. Our model is able to catch some additional scenarios where officers have little to no discretion in initiation such as when they are flagged down by a victim while on patrol or directly observe an assault or battery (or some other serious incident).[34] We highlight this as a way to underscore the value of the ML model in understanding the level of discretion a police officer has in initiation beyond factors that could be easily identified in the existing administrative data, such as if a call for service was placed.

[34]Other instances include when police officers have active search warrants or when people are positively ID'ed for an offense by a civilian.

7.2.1 Alternative classifications for discretion

Studies that seek to address the use of arrests as a proxy for criminal behavior often rely on existing heuristics such as the index crime classification developed by the FBI Uniform Crime Report (UCR)(Rivera, 2022; Lum and Nagin, 2017). Part I or index crimes comprise of the most serious crimes, mostly felonies with identifiable victims and are often low-discretionary.[35] Part II offenses or non-index crimes are often misdemeanors with no identifiable victim and typically come with higher police discretion in enforcement.[36] To underscore the value of our ML classification method, we compare classifications of discretion using existing methods in the literature. Column 3 in Table V displays this non-index/high discretionary categorization with the FBI category identified as well. Some researchers improve on this non-index classification by excluding simple assault and simple battery from the high discretionary classification (non-index crimes) (shown in column 4 in Table V. We see that using both of these index measures of discretion gives a very different takeaway in terms of behavior C2C youth engage in, with the vast majority of arrests in our outcome sample classified as non-index/high discretion (57-68%).

Lastly, while our ML classification provides variation *within charges*, we could also use our ML classification to indicate which categories of arrests are more or less likely to be high discretion. For example, our ML model predicts that 71% of all arrests related to Chicago Municipal Code violations are high discretion. If other researchers wanted to use the takeaways from our ML classification and did not have access to individual arrest narratives, they could classify all arrests within this charge as high discretion using a majority classification cutoff rule (i.e. if more than 50% of a charge is classified as high discretion, we consider all the arrests in that charge category as high discretion). Column 5 in Table V shows this classification method with 28% overall classified as high discretion. This is closer to our individual-arrest ML method that finds 33% of all arrests are high-discretion.

7.3 Impact of C2C on Discretionary Arrests

Using our new classification of arrests, we can now turn back to our RCT framework to understand what types of arrests C2C impacted as it pertains to low or high discretion. Table

[35]Index crimes comprise of non-negligent homicide, sexual assault, robbery, aggravated assault, aggravated battery, burglary, larceny, motor vehicle theft.

[36]Part II offenses include other assaults (simple assault and simple battery), forgery and counterfeiting, fraud, embezzlement, buying, receiving and possessing stolen property, vandalism, carrying and possessing weapons, prostitution and commercialized vice, sex offenses (except forcible rape and prostitution), drug sales, possession and use, gambling, offenses against the family and children, driving under the influence, liquor laws, drunkenness, disorderly conduct, vagrancy, all other offenses (except traffic), suspicion, curfew and loitering law violations, and runaways.

VI presents the extensive and intensive margin results for the Intent-to-Treat (ITT) effects on arrests broken down by those that were initiated with high discretion from the officer and those that were not over time using our ML method and compared to other existing methods of discretion classification. Using our ML classification method (last column in Table VI), we find that the effect of the C2C program on reducing the likelihood of having any arrest after randomization seems to be driven primarily by a reduction in arrests where officers have little discretion in initiation. The estimated effect on low-discretionary arrests are large (a 5.5 percentage point decrease, with 30.6% control mean in 24 months post randomization) and similar in magnitude to the overall estimated extensive margin effects on any arrest and statistically significant at the 1% level[37]. By contrast, there is little evidence that the program is substantively decreasing arrests where the initial youth/police contact was initiated at the discretion of the officer. The high discretionary arrest estimate is small, negative and not statistically significant at conventional levels in almost every outcome period. However, looking at the 95% confidence interval we cannot rule out effects that may be substantively meaningful.

Looking at the intensive margin effects, the pattern of the estimated effects follows the same pattern as that seen in the extensive margin, where the coefficients for arrest initiated with little to no officer discretion are larger and statistically significant compared to those for arrests initiated at the discretion of an officer (see Table VI for details). The high discretionary arrest estimates are close to zero, while low discretionary arrest estimates are much larger and negative. In fact, the 12 month post randomization estimates for low discretionary arrests are significant at the 5 percent level, highlighting that measuring the discretionary arrests may have removed some of the noise we were seeing in the overall number of arrest results.

Lastly, we also compare the RCT results using our ML classification method with other existing methods in the literature (non-index or index classification). Using the index classification method, we would have concluded the program is mainly reducing *high discretion* arrests. When we include simple assault and battery in the low-discretion/index bucket, the results show mixed findings: reduction on both high and low margins of discretion. It is not until we use the majority classification method from our ML model do we begin to see results that mimic our individual arrest ML classification model (last column of results). These findings suggest the value of our data-driven approach to account for police discretion and helps us understand where changes in behavior are actually coming from. Given the within charge variation that we identified in our data, our individual ML method is our

[37]These RCT results are not sensitive to key ML tuning parameters that we discuss in further detail in the Appendix.

preferred specification. However, we also see the value of using our data-driven approach to inform which categories of arrests are more or less likely to be associated with discretion if researchers do not have access to individual arrest information (for example, (Agan et al., 2023) cited and used our ML classification method in this manner).

7.4 Robustness Checks

7.4.1 Discretionary classification within charge level categories

We find that the C2C program does not significantly decrease the arrests that result from a discretionary initiation by a police officer. These arrests are concentrated in possession of illegal substances, gambling, Chicago municipal violations, traffic infractions, and weapons violations, among others. One interpretation of these findings could be that a program like C2C could not change the criminal behavior underlying more discretionary initiation arrests. Another interpretation could be that because of the high level of contact this population has with police, police may be more likely to observe and then enforce this type of arrest regardless of underlying behavior change.

This paper will not be able to discern between these interpretations or others. However, we do look to see within charge level classification, for those where we actually see some substantial variation in discretion (e.g. the other or drug category), what are the C2C impacts (see Table A.X in Appendix). We want to highlight that we are cautious about interpreting these results given the power concerns once we begin to split the data in these small samples. However, this exploratory analysis does suggest that within the drug categories, the positive (adverse) ITT estimate was being driven by high discretionary arrests, and C2C may be decreasing the low-discretionary drug arrests. For the "other" arrest category, we see that the estimates are more precise for the low-discretionary category compared to the overall estimates (Table III, suggesting that the discretionary classification is removing some noise in impact (about 47% of the "other" arrests were classified as high discretionary).[38]

7.4.2 Another Form of Discretion: Police Stops

As an additional robustness check, the second part of our discretionary criminal justice contact measurement exercise focuses on using an alternative measure of police contact that relies on discretion: police stops. The last few decades have seen a steep rise in the use of proactive police strategies, which intentionally raise the frequency of police contact through street stops and field interrogations. These strategies have also been heavily criticized for

[38]Other categories of arrests can be provided upon ask, but due to space limitations were left out of the Appendix.

targeting minority and young populations in specific neighborhoods (Gelman et al., 2007; Stoudt et al., 2011; National Academies of Sciences et al., 2018). [39] To our knowledge, we are unaware of any program evaluation for individuals that look at street stops as a criminal justice outcome, despite street stops being the most common interaction individuals have with law enforcement and their documented negative impact on health, mental well-being, educational outcomes, fear of police, and political engagement, particularly among communities of color (Weitzer et al., 2008; Butler, 2014; Bandes et al., 2019; Futterman et al., 2016; Geller, 2019; Bell, 2017; Weaver and Geller, 2019; Del Toro et al., 2019; Pickett et al., 2021; Bacher-Hicks and de la Campa, 2020).

The legal rules governing stop and frisk permit officers to have a wide degree of discretion in who they can stop and under what circumstances. The standard set from the 1968 Supreme Court Terry v. Ohio (1968) involved a pedestrian stop that defined the parameters around "reasonable suspicion." The court ruled that it is not a violation of the Fourth Amendment's prohibition on unreasonable searches and seizures when a police officer stops a suspect on the street and questions them without probable cause to arrest, so long as the police officer has a reasonable suspicion that the person has committed, is committing, or is about to commit a crime.[40]

In Chicago, police officers are supposed to fill out a stop record (called investigatory stop report (ISR) at CPD) anytime a contact has officially become a stop: when someone is being detained and they can't walk away voluntary. This is different from a field investigation where police officers can ask people questions, but the civilian is free to go at anytime. Documentation from CPD states that police officers are supposed to make the person aware when an encounter is a field investigation (a person doesn't have to answer and they are free to go) vs when something is an official investigatory stop (the police officer suspects

[39] In Chicago, the ACLU came to an agreement with CPD to institute new stop regulations in 2016 after they found that CPD engaged in a pattern of unconstitutional street stops. The newly instituted street stop practices and procedures led to a substantial reduction in street stops coupled with a corresponding increase in vehicle stops (Hausman and Kronick, 2021). Given this behavioral change, we will investigate C2C's impact on both pedestrian and vehicle stops given the common use of pretext traffic stops as a policing strategy in Chicago. Only a small portion of our outcome period for this study happens prior to 2016 (November 2015-December 2015), where 18% of our study sample had been randomized. The vast majority of our study sample and study period for stops will cover the time period after the stops policy change.

[40] Several subsequent court decisions after Terry v. Ohio, the concept of "reasonable suspicion" was expanded to include location as well as behavior. For instance, in the Supreme Court case of Illinois v. Wardlow (2000), a person's presence in a "high-crime area" can be relevant in determining whether a person's individual behavior is sufficiently suspicious. This court case is key to the detainment of youth who flee from officers in high crime areas, a very common scenario on the streets of Chicago. The courts have also determined within the police discretion to make a warrantless custodial arrest for a very minor offense, such as a seat belt violation, that is punishable only by a fine (Atwater v. Lago Vista, 2001).

a crime is happening, about to happen, or has happened). [41]. Officer-initiated pedestrian and vehicle stops are often sensitive to officer effort and discretion as officers choose whether to investigate and/or intervene. Given that police stops are a measure of justice contact with a high degree of discretion, we believe this is an important juvenile justice outcome to understand as it is often the first documented point of contact between police and civilians.

7.4.3 Program Effects on Police Stops

We examine the effect of the C2C program on the likelihood of being stopped by the police, regardless of whether the stop culminates in an arrest (Table A.XI). It is notable that there this is a high level of interaction between the police and study youth following randomization. Within just six months of randomization, over a fifth of the youth have been stopped by a police officer. Within a year after randomization, almost a third of youth have been stopped by a police officer. However, the program appears to have no discernible impact on the likelihood that a young person experiences a stop. Twenty four months post randomization, when almost half of the control group have been stopped, we do find a small adverse impact with C2C youth slightly more likely to be stopped by a police officer (ITT effect of 3.6 percentage points, with p-value < 0.10, representing an 8% increase in likelihood of being stopped). This could be a result of C2C youth spending more time in new neighborhoods of Chicago and generally participating in more activities outside the house, something we know the program encourages. It is also possible that due to the program's effects, C2C creates an incapacitation effect with control youth being incarcerated and therefore are less likely to be stopped. Looking at incarceration as an outcome, we do find negative treatment effects however these effects are not statistically significant at conventional levels (see Table A.III and section A.1.7 in Appendix). We do not observe an adverse stop impact in other time periods, among the intensive stop margin, or when we break this out by stop type.

While many of these point estimates are noisy, we are able to rule out with the extensive margin estimates that C2C leads to a reduction in the likelihood of being stopped. [42] Here, looking at the 95% confidence interval of any stop within 36 months of randomization, the longest run impact estimate for which we have for full sample, we are able to rule out ITT program reduction effects that are as small as 5 percent of the control mean.

[41] For more details on the CPD documentation required for both street and traffic stops, please see section A.1.6.2 in the Appendix

[42] The precision with which we are able to measure the effect of the program on stops is greater than that with which we are able to measure the effect of the program on arrests that are initiated as a result of officer discretion.

7.5 Heterogeneity analysis

Lastly, we explore whether or not the program had heterogeneous treatment effects. These results should be interpreted with caution, given we were not powered to detect subgroup effects and there are a large number of hypothesis tests involved. We look to see if the main RCT effects had differential impacts by three different subgroups: youth with and without a prior arrest, youth with and without a prior arrest for a violent offense, youth with and without a prior gap in school enrollment in Table A.XII in the Appendix. We focus on reporting the ITT effects and our main outcome window: 24 months post randomization.

Focusing on the violent-crime arrest outcomes first, we find suggestive evidence the program is more effective for those with prior criminal justice contact. The negative point estimates on the interaction term suggest that the program is more successful at reducing arrests for violent offenses among youth who may be considered higher need – youth with a baseline arrest, youth with a baseline violent-crime arrest, or youth with a school enrollment gap. However, the interaction term is not significantly different from zero, and therefore we cannot draw any definitive conclusions.

In contrast, when we look at all arrests, we find suggestive evidence that the program is more successful at preventing arrests for youth without a baseline arrest, a baseline violent-crime arrest, or without a school enrollment gap. However, in most instances the interaction is too imprecisely estimated to tell a clear story. The one exception is when we look at the impact of the program on number of arrests. As a reminder, we did not detect a significant effect on the number of arrests when using the full study sample. When we look at the subgroup effects, it appears this null overall finding is masking a large negative point estimate (-0.165, p-value < 0.001) among youth without a prior arrest/prior arrest for violent-crime, and a positive large point estimate for youth with a prior arrest/prior arrest for violent-crime (0.506, p-value < 0.05). To better understand what might be driving this potential adverse impact for youth with prior criminal justice contact, we break out arrests into high/low discretion in Table A.XIII. We see a large and positive point estimate on the interaction term for high discretion arrests among youth with a baseline arrest for a violent-crime. These findings suggest that young people who are known to police through previous serious incidents may be more subject to discretionary contact, and the program may be creating more opportunities for contact through the activities in the neighborhood. This is consistent with the literature in this area that highlights previous criminal justice contact is an important factor contributing to police discretion (Carrington et al., 2003; McGlynn-Wright et al., 2022).

These results suggest the program may have differential impacts depending on the youth served and the outcome of interest. It also helps us better understand why we saw a null

effect on number of arrests for the full study sample, and the value-add of disaggregating arrests by discretion. However, given most of the estimates are imprecise and the multiple tests involved, we cannot make any strong conclusions regarding heterogeneous treatment effects.

8 Discussion and Next Steps

Taken as a whole, these results suggest that C2C leads to a sustained change in youth behavior and a reduction in violence. We see significant reductions in low-discretion arrests. Importantly, we see these reductions are being driven by violent offenses as well as other offenses related to interpersonal conflict (e.g. disorderly conduct, etc) that are flagged to police through calls for service or citizen complaints. In fact, as Table V notes, violent arrests are only about a third of all low discretionary arrests, highlighting the value of moving beyond a charge level analysis to understand the other types of behavior C2C is impacting. Given the little causal evidence that exists around social programming that can reduce and sustain impacts on violence, these findings are noteworthy. Importantly, our study follows most youth into adulthood, highlighting a program that can have longer term impacts.

We explore why C2C reduced violent-offense arrests and other low discretionary arrests that are often associated with interpersonal conflict in our qualitative work where we conducted over 12 focus groups with 69 C2C participants and 19 interviews with C2C staff. Some of the key themes that came out of these data suggest that the program leads to increased future orientation, and increased ability to process trauma, an increase in self-esteem and self-worth, an increase in social and emotional skills, and new skills to manage stressful situations. Focus group respondents also stressed the importance of strength based mentoring from an adult. For example, one youth said, "He'd been there, done that, he'd been my age before. ... actually, listening and taking advice from someone who knows[s] and not just trying to put off their opinion on you, someone who was actually in this situation and overcame it." Trauma-informed cognitive behavioral therapy can help youth build a new set of coping skills and tools to manage the difficult environments youth are too often subject to in neighborhoods in Chicago. For example, one youth said "I learned how to walk away and not act on everything and make a permanent decision on a temporary situation." Furthermore, one C2C therapist highlighted, the shift in perspective from "what is wrong with you?" to asking "what happened to you" can be a beneficial perspective and reframe existing challenges in a trauma-informed lens that puts healing at the forefront. The consistent, positive adult support and the new experiences youth have in the program can also serve as pathways to restore physical safety. Research has highlighted that social support is

one of the most powerful protections against becoming overwhelmed with stress and trauma (Van der Kolk, 2015). The positive experiences and extensive time with the mentor/advocate served as opportunities for new coping skills to be modeled and practiced. The existing psychology literature has highlighted that the more frequently a person performs a behavior or skill, the more habitual and automatic it becomes, which can help youth respond in new moments of stress (Van der Kolk, 2015; Wilson, 2004; Beck and Beck, 2011). The program model allows for opportunities to practice the new skills youth learn in C2C, suggesting this is one potential mechanism for why we see the sustained program impacts. Given the high levels of trauma and adverse experience exposure among youth in the program, our findings highlight the importance of trauma-informed programming in supporting youth's ability to safely navigate their environment.[43]

The discretionary classification of our arrests enables us to confirm that these are true changes in civilian behavior and are not driven by youth avoiding police given the lack of police discretion in the arrests that were most impacted by the program. Likewise, while some officers in CPD and CPD leadership were aware this program existed, there were no directives to officers to treat these youth differently and no way to distinguish non-C2C youth from C2C youth in the community. The arrest classification methodology we developed in this paper helps us better understand behavior change when using arrests as a proxy for criminal behavior and could potentially be used by other researchers seeking to incorporate the role of police discretion in program evaluations.

Finally, we pursued the discretionary classification exercise in an effort to better understand how the C2C program was affecting behavior. However, we do think it is worth further exploring the context, if any, in which discretionary contact with police may be reduced. Our paper can not speak to the tradeoffs or the benefits or costs of various police strategies that lead to high contact with police. However, our paper does highlight a potential collateral effect of these policies by demonstrating a situation where youth behavior on more serious dimensions changes (including a reduction in violence), and yet this did not impact high discretionary arrests consistently or the propensity of being stopped by police. By incorporating the role of police behavior in evaluations of programs, we believe this line of research can help with understanding why we may or may not see impacts with individual programs that moves beyond the role the individual plays and begins to incorporate the policing systems around youth. Given the on-going policy debate around police reform and the need to maintain public safety, our paper provides an empirical approach to connect these two

[43]On average, youth participants had experienced 7 traumatic or adverse experiences prior to the program starting, and almost all had experienced at least one.

conversations by beginning to distinguish between police behavior and civilian behavior in measures of justice system contact.

Ensuring high-need youth have access to mental and behavioral health services may help youth build the social-emotional skills they need to succeed later in life as well. Our future work in this project aims to produce a complementary paper that will cover the full educational outcomes, including graduation from high school, once data becomes available. We also plan to include other longer term outcomes such as incarceration and conduct a full cost-benefit analysis for this program.

9 Conclusion

This paper finds that a youth behavioral health intervention in Chicago, C2C, can effectively engage harder to reach youth, substantially reduce violence involvement, and reduce the probability of arrest in the short and longer term. Using a new method of classifying arrests, we find that looking at the charges of arrests can obscure heterogeneity with respect to the level of discretion involved in initiation from a police officer. Through the use of machine learning methods, we find that C2C reductions in arrests are concentrated among arrests with little discretion in police initiation, with no detectable impact on more discretionary police contact.

In the context of other successful CBT-based programs in Chicago, C2C highlights the possibility of engaging a harder-to-reach population of youth outside of school without having to wait until someone reaches adulthood or are detained (Heller et al., 2017; Bhatt et al., 2023). We believe these results highlight the promise of behavioral health interventions in developing meaningful and sustained impacts for youth by providing them with tools to navigate challenging environments that are often plagued with a lack of safety and economic opportunity.

10 References

Agan, Amanda, Jennifer L Doleac, and Anna Harvey (2023) "Misdemeanor prosecution," *The Quarterly Journal of Economics*, 138 (3), 1453–1505.

Aizer, Anna and Joseph J Doyle (2015) "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges," *The Quarterly Journal of Economics*, 130 (2), 759–803.

Allen, Terrence T (2005) "Taking a juvenile into custody: Situational factors that influence police officers' decisions," *J. Soc. & Soc. Welfare*, 32, 121.

Ang, Desmond (2021) "The effects of police violence on inner-city students," *The Quarterly Journal of Economics*, 136 (1), 115–168.

Ang, Desmond and Jonathan Tebes (2020) "Civic Responses to Police Violence," *Harvard Kennedy School, John F. Kennedy School of Government*.

Angrist, Joshua D and Jörn-Steffen Pischke (2008) *Mostly harmless econometrics*: Princeton university press.

Antonovics, Kate and Brian G Knight (2009) "A new look at racial profiling: Evidence from the Boston Police Department," *The Review of Economics and Statistics*, 91 (1), 163–177.

Arbour, William (2021) *Can Recidivism be Prevented from Behind Bars?: Evidence from a Behavioral Program*: University of Toronto, Department of Economics 4060 1 Online-Ressource.

Athey, Susan and Guido W Imbens (2019) "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Ba, Bocar A, Dean Knox, Jonathan Mummolo, and Roman Rivera (2021) "The role of officer race and gender in police-civilian interactions in Chicago," *Science*, 371 (6530), 696–702.

Ba, Bocar, Patrick Bayer, Nayoung Rim, Roman Rivera, and Modibo Sidibé (2021) "Police Officer Assignment and Neighborhood Crime,"Technical report, National Bureau of Economic Research.

Bacher-Hicks, Andrew and Elijah de la Campa (2020) "Social Costs of Proactive Policing: The Impact of NYC's Stop and Frisk Program on Educational Attainment,"Technical report, Working paper.

Bandes, Susan A, Marie Pryor, Erin M Kerrison, and Phillip Atiba Goff (2019) "The mismeasure of Terry stops: Assessing the psychological and emotional harms of stop and frisk to individuals and communities," *Behavioral sciences & the law*, 37 (2), 176–194.

Barnes, Geoffrey C, Jordan M Hyatt, and Lawrence W Sherman (2017) "Even a little bit helps: An implementation and experimental evaluation of cognitive-behavioral therapy for high-risk probationers," *Criminal Justice and Behavior*, 44 (4), 611–630.

Beck, JS and AT Beck (2011) "Cognitive behavior therapy: basics and beyond, vol. 2."

Bell, Monica C (2017) "Police reform and the dismantling of legal estrangement," *The Yale Law Journal*, 2054–2150.

Bhatt, Monica P, Sara B Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman (2023) "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago*," *The Quarterly Journal of Economics*, qjad031, 10.1093/qje/qjad031.

Blattman, Christopher, Sebastian Chaskel, Julian C Jamison, and Margaret Sheridan (2023) "Cognitive Behavioral Therapy Reduces Crime and Violence over Ten Years: Experimental Evidence," *American Economic Review: Insights*, 5 (4), 527–545.

Blattman, Christopher, Julian C Jamison, and Margaret Sheridan (2017) "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia," *American Economic Review*, 107 (4), 1165–1206.

Blattman, Christopher, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues, and Margaret Sheridan (2016) "Measuring the measurement error: A method to qualitatively validate survey data," *Journal of Development Economics*, 120, 99–112.

Borduin, Charles M, Barton J Mann, Lynn T Cone, Scott W Henggeler, Bethany R Fucci, David M Blaske, and Robert A Williams (1995) "Multisystemic treatment of serious juvenile offenders: long-term prevention of criminality and violence.," *Journal of consulting and clinical psychology*, 63 (4), 569.

Bradley, Andrew P (1997) "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, 30 (7), 1145–1159.

Breiman, Leo (2001) "Random forests," *Machine learning*, 45 (1), 5–32.

Breiman, Leo, Jerome Friedman, Richard A Olshen, and Charles J Stone (1984) "Classification and regression trees Chapman & Hall," *New York*.

Brown, Robert A, Kenneth J Novak, and James Frank (2009) "Identifying variation in police officer behavior between juveniles and adults," *Journal of criminal justice*, 37 (2), 200–208.

Butler, Paul (2014) "Stop and frisk and torture-lite: police terror of minority communities," *Ohio St. J. Crim. L.*, 12, 57.

Carrington, Peter J, Jennifer L Schulenberg, Anne Brunelle, Joanna Jacob, and Ian Pickles (2003) *Police discretion with young offenders*: Department of Justice Canada Ottawa.

Chalfin, Aaron (2015) "Economic costs of crime," *The encyclopedia of crime and punishment*, 1–12.

Chalfin, Aaron and Felipe Goncalves (2020) "The pro-social motivations of police officers,"Technical report, Working Paper.

Chandler, Dana, Steven D Levitt, and John A List (2011) "Predicting and preventing shootings among at-risk youth," *American Economic Review*, 101 (3), 288–292.

Cho, Sungwoo, Felipe Gonçalves, and Emily Weisburst (2021) "Do Police Make Too Many Arrests?".

Cloitre, Marylene, Bradley C Stolbach, Judith L Herman, Bessel van der Kolk, Robert Pynoos, Jing Wang, and Eva Petkova (2009) "A developmental approach to complex PTSD: Childhood and adult cumulative trauma as predictors of symptom complexity," *Journal of traumatic stress*, 22 (5), 399–408.

Council, National Research et al. (2011) "Social and Economic Costs of Violence: Workshop Summary."

Cunha, Flavio, James J Heckman, Lance Lochner, and Dimitriy V Masterov (2006) "Interpreting the evidence on life cycle skill formation," *Handbook of the Economics of Education*, 1, 697–812.

Cunha, Flavio, James J Heckman, and Susanne M Schennach (2010) "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78 (3), 883–931.

D'Agostino, Jerome V, Emily Rodgers, and Susan Mauck (2018) "Addressing inadequacies of the observation survey of early literacy achievement," *Reading Research Quarterly*, 53 (1), 51–69.

Davis, Jonathan MV and Sara B Heller (2020) "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs," *Review of economics and statistics*, 102 (4), 664–677.

De Rosa, R, D Pelcovitz, J Rathus et al. (2004) "SPARCS (Structured Psychotherapy for Adolescents Responding to Chronic Stress)."

Del Toro, Juan, Tracey Lloyd, Kim S Buchanan et al. (2019) "The criminogenic and psychological effects of police stops on adolescent black and Latino boys," *Proceedings of the National Academy of Sciences*, 116 (17), 8261–8268.

Desmond, Matthew, Andrew V Papachristos, and David S Kirk (2016) "Police violence and citizen crime reporting in the black community," *American sociological review*, 81 (5), 857–876.

Dinarte-Diaz, Lelys and Pablo Egana-delSol (2024) "Preventing violence in the most violent contexts: Behavioral and neurophysiological evidence from el salvador," *Journal of the European Economic Association*, 22 (3), 1367–1406.

Doleac, Jennifer L (2020) "Encouraging desistance from crime," *Available at SSRN*.

Dube, Oeindrila, Sandy Jo MacArthur, and Anuj K Shah (2023) "A cognitive view of policing,"Technical report, National Bureau of Economic Research.

Dynarski, Mark, Philip M Gleason, Anu Rangarajan, and Robert G Wood (1998) "Impacts of dropout prevention programs,"Technical report, Mathematica Policy Research.

Feigenberg, Benjamin and Conrad Miller (2022) "Would eliminating racial disparities in motor vehicle searches have efficiency costs?" *The Quarterly Journal of Economics*, 137 (1), 49–113.

Flannery, Daniel J, Kelly L Wester, and Mark I Singer (2004) "Impact of exposure to violence in school on child and adolescent mental health and behavior," *Journal of community psychology*, 32 (5), 559–573.

Force, Police Accountability Task (2016) "Recommendations for reform: restoring trust between the Chicago Police and the communities they serve," *Chicago, IL: Chicago Police Accountability Task Force*.

Ford, Julian D, J Kirk Hartman, Josephine Hawke, and John F Chapman (2008) "Traumatic victimization, posttraumatic stress disorder, suicidal ideation, and substance abuse risk among juvenile justice-involved youth," *Journal of Child & Adolescent Trauma*, 1 (1), 75–92.

Forman Jr, James (2017) *Locking up our own: Crime and punishment in Black America*: Farrar, Straus and Giroux.

Futterman, Craig B, Chaclyn Hunt, and Jamie Kalven (2016) "Youth/police encounters on Chicago's south side: Acknowledging the realities," *U. Chi. Legal F.*, 125.

Geller, Amanda (2019) "Policing America's children: Police contact among urban teens," *Unpublished Manuscript. Fragile Families Working Paper WP18-02-FF*.

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss (2007) "An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias," *Journal of the American statistical association*, 102 (479), 813–823.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019) "Text as data," *Journal of Economic Literature*, 57 (3), 535–74.

Gobbo, Andre (2023) "The economic costs of gun violence in the United States."

Goel, Sharad, Justin M Rao, and Ravi Shroff (2016) "Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy," *The Annals of Applied Statistics*, 10 (1), 365–394.

Goldstein, Herman (1963) "Police discretion: The ideal versus the real," *Public Administration Review*, 140–148.

Gonçalves, Felipe M and Steven Mello (2023) "Police discretion and public safety,"Technical report, National Bureau of Economic Research.

Goncalves, Felipe and Steven Mello (2021) "A few bad apples? Racial bias in policing," *American Economic Review*, 111 (5), 1406–41.

Hand, David J (2009) "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine learning*, 77 (1), 103–123.

Hausman, David and Dorothy Kronick (2021) "When Police Sabotage Reform by Switching Tactics," *Available at SSRN 3192908*.

Heckman, James J and Tim Kautz (2013) "Fostering and measuring skills: Interventions that improve character and cognition."

Heller, Sara B (2014) "Summer jobs reduce violence among disadvantaged youth," *Science*, 346 (6214), 1219–1223.

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack (2017) "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago," *The Quarterly Journal of Economics*, 132 (1), 1–54.

Herrera, Carla, David L DuBois, and Jean Baldwin Grossman (2013) "The role of risk: mentoring experiences and outcomes for youth with varying risk profiles.," *MDRC*.

Hinton, Elizabeth and DeAnza Cook (2021) "The mass criminalization of Black Americans: A historical overview," *Annual Review of Criminology*, 4, 261–286.

Hjalmarsson, Randi (2008) "Criminal justice involvement and high school completion," *Journal of Urban Economics*, 63 (2), 613–630.

Hoekstra, Mark and CarlyWill Sloan (2020) "Does race matter for police use of force? Evidence from 911 calls,"Technical report, National Bureau of Economic Research.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013) "An introduction to statistical learning (Vol. 112, p. 18)."

of Justice, Department (2017) "Report on the Investigation of the Chicago Police Department,"Technical report, `https://www.justice.gov/opa/file/925846/download`.

Katz, Lawrence F, Jeffrey R Kling, and Jeffrey B Liebman (2001) "Moving to opportunity in Boston: Early results of a randomized mobility experiment," *The Quarterly Journal of Economics*, 116 (2), 607–654.

Kelling, George L (1999) *Broken windows and police discretion*: US Department of Justice, Office of Justice Programs, National Institute of Justice.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007) "Experimental analysis of neighborhood effects," *Econometrica*, 75 (1), 83–119.

Kling, Jeffrey R, Jens Ludwig, and Lawrence F Katz (2005) "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," *The Quarterly Journal of Economics*, 120 (1), 87–130.

Van der Kolk, Bessel A (2015) *The body keeps the score: Brain, mind, and body in the healing of trauma*: Penguin Books.

Landenberger, Nana A and Mark W Lipsey (2005) "The positive effects of cognitive–behavioral programs for offenders: A meta-analysis of factors associated with effective treatment," *Journal of experimental criminology*, 1 (4), 451–476.

Larson, Katherine A and Russell W Rumberger (1995) "ALAS: Achievement for Latinos through academic success," *Staying in School. A Technical Report of Three Dropout Prevention Projects for Junior High School Students with Learning and Emotional Disabilities*.

Legewie, Joscha and Jeffrey Fagan (2019) "Aggressive policing and the educational performance of minority youth," *American Sociological Review*, 84 (2), 220–247.

Lerman, Amy E and Vesla M Weaver (2014) *Arresting citizenship*: University of Chicago Press.

Linn, Edith (2009) *Arrest decisions: What works for the officer?* (5): Peter Lang.

Lum, Cynthia and Daniel S Nagin (2017) "Reinventing american policing," *Crime and justice*, 46 (1), 339–393.

Maclin, Tracey and Maria Saverese (2018) "Martin Luther King, Jr. and Pretext Stops (and Arrests): Reflections on How Far We Have Not Come Fifty Years Later," *U. Mem. L. Rev.*, 49, 43.

Margolin, Gayla and Elana B Gordis (2000) "The effects of family and community violence on children," *Annual review of psychology*, 51 (1), 445–479.

Mastrofski, Stephen D (2004) "Controlling street-level police discretion," *The annals of the American academy of political and social science*, 593 (1), 100–118.

McGlynn-Wright, Anne, Robert D Crutchfield, Martie L Skinner, and Kevin P Haggerty (2022) "The usual, racialized, suspects: The consequence of police contacts with black and white youth on adult arrest," *Social problems*, 69 (2), 299–315.

Mendias, Claudia and E James Kehoe (2006) "Engagement of policing ideals and their relationship to the exercise of discretionary powers," *Criminal Justice and Behavior*, 33 (1), 70–92.

Mullainathan, Sendhil and Jann Spiess (2017) "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 31 (2), 87–106.

Myers, Stephanie M (2002) *Police encounters with juvenile suspects: Explaining the use of authority and provision of support*: State University of New York at Albany.

Natapoff, Alexandra (2017) "Misdemeanors," *University of California, Irvine School of Law: Legal Studies Research Paper Series*.

Nickels, Ernest L (2007) "A note on the status of discretion in police research," *Journal of Criminal Justice*, 35 (5), 570–578.

Owens, Emily G (2017) "Testing the school-to-prison pipeline," *Journal of Policy Analysis and Management*, 36 (1), 11–37.

Pickett, Justin, Amanda Graham, and Frank Cullen (2021) "The American Racial Divide in Fear of the Police."

Rivera, Roman (2022) "The effect of minority peers on future arrest quantity and quality,"Technical report, Technical report.

Rosenfeld, Richard, Jeff Rojek, and Scott Decker (2012) "Age matters: Race differences in police searches of young and older male drivers," *Journal of research in crime and delinquency*, 49 (1), 31–55.

Saito, Takaya and Marc Rehmsmeier (2015) "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, 10 (3), e0118432.

Santiago, Catherine DeCarlo, Tali Raviv, and Lisa H Jaycox (2018) *Creating healing school communities: School-based interventions for students exposed to trauma.*: American Psychological Association.

National Academies of Sciences, Engineering, Medicine et al. (2018) *Proactive policing: Effects on crime and communities*: National Academies Press.

Sealock, Miriam D and Sally S Simpson (1998) "Unraveling bias in arrest decisions: The role of juvenile offender type-scripts," *Justice Quarterly*, 15 (3), 427–457.

Skogan, Wesley G and Kathleen Frydl (2004) "Fairness and effectiveness in policing: The evidence."

Smith, Douglas A and Christy A Visher (1981) "Street-level justice: Situational determinants of police arrest decisions," *Social problems*, 29 (2), 167–177.

Stoudt, Brett G, Michelle Fine, and Madeline Fox (2011) "Growing up policed in the age of aggressive policing policies," *NYL Sch. L. Rev.*, 56, 1331.

Stroshine, Meghan, Geoffrey Alpert, and Roger Dunham (2008) "The influence of "working rules" on police suspicion and discretionary decision making," *Police quarterly*, 11 (3), 315–337.

Stuart, Forrest (2016) *Down, out, and under arrest: Policing and everyday life in skid row*: University of Chicago Press.

Van Dijk, Sheri (2013) *DBT made simple: A step-by-step guide to dialectical behavior therapy*: New Harbinger Publications.

Weaver, Vesla M and Amanda Geller (2019) "De-policing America's youth: Disrupting criminal justice policy feedbacks that distort power and derail prospects," *The ANNALS of the American Academy of Political and Social Science*, 685 (1), 190–226.

Weaver, Vesla M, Andrew Papachristos, and Michael Zanger-Tishler (2019) "The great decoupling: The disconnection between criminal offending and experience of arrest across two cohorts," *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 5 (1), 89–123.

Weisburst, Emily K (2022) ""Whose help is on the way?": The importance of individual police officers in law enforcement outcomes," *Journal of Human Resources*.

Weitzer, Ronald, Steven A Tuch, and Wesley G Skogan (2008) "Police–community relations in a majority-Black city," *Journal of Research in Crime and Delinquency*, 45 (4), 398–428.

Western, Bruce (2006) *Punishment and inequality in America*: Russell Sage Foundation.

Westfall, Peter H and S Stanley Young (1993) *Resampling-based multiple testing: Examples and methods for p-value adjustment*, 279: John Wiley & Sons.

Willis, James J (2013) "First-line supervision and strategic decision making under compstat and community policing," *Criminal Justice Policy Review*, 24 (2), 235–256.

Willis, James J and Stephen D Mastrofski (2017) "Understanding the culture of craft: lessons from two police agencies," *Journal of crime and justice*, 40 (1), 84–100.

Wilson, Timothy D (2004) *Strangers to ourselves*: Harvard University Press.

Wu, Xiaoyun and Cynthia Lum (2017) "Measuring the spatial and temporal patterns of police proactivity," *Journal of quantitative criminology*, 33 (4), 915–934.

11 Tables and Figures

Table I: C2C Study Youth Baseline Characteristics

Variable	Mean T	Mean C	Difference
Demographics			
Age	15.79	15.77	0.02
% Black	0.94	0.96	-0.01
% Hispanic	0.05	0.04	0.01
% Female	0.43	0.41	0.02
Number of arrests			
Number of prior arrests	1.41	1.57	-0.16
Number of prior violent arrests	0.38	0.45	-0.07*
Number of prior property arrests	0.36	0.36	0.00
Number of prior drug arrest	0.10	0.10	0.00
Number of prior other arrests	0.57	0.67	-0.09
Any arrest			
Any prior arrest	0.35	0.36	-0.01
Any prior violent arrest	0.22	0.23	-0.01
Any prior property arrest	0.17	0.18	-0.02
Any prior drug arrest	0.06	0.05	0.01
Any prior other arrest	0.21	0.23	-0.03*
Victimizations			
Number of prior victimizations	0.90	0.94	-0.04
Any prior victimization	0.49	0.48	0.00
Education			
Had free/reduce lunch status	0.96	0.95	0.01
School grade at baseline	9.78	9.72	0.06
Had an enrollment gap	0.20	0.20	-0.01
Observations	1052	1022	
P-value on Join F Test	0.58		

Notes: Mean differences for the treatment vs. control groups was estimated using a linear regression with referral level and cohort fixed effects (randomization was conducted within referral source and cohort). Standard errors are robust. Stars indicator the following p-values: *** p<0.01, ** p<0.05, * p<0.1. School grade at baseline is only available for 1,996 youth. Variables included in the F-test are all demographics, indicators for every type of prior arrest, indicator for any prior victimization, indicator for prior free/reduced lunch status, and prior disconnection from school.

Table II: Trauma-informed CBT sessions & Mentoring hours, All Cohorts

			SPARCS Hours				Mentoring Hours				Total Hours
	N Treat	Take-up Rate	25th Percentile	50th Percentile	75th Percentile	Average	25th Percentile	50th Percentile	75th Percentile	Average	Average
All Participants	1052	62%	4	9	14	9	126	177	223	177	186
Gap in Enrollment											
Baseline gap in enrollment	206	45%	2	10	13	8	119	163	207	167	175
No baseline gap in enrollment	846	66%	4	9	14	10	126	182	226	178	188
Any Baseline Arrest											
Has prior arrest	363	51%	3	8	14	9	124	178	224	183	192
No prior arrest	689	68%	5	10	14	10	126	177	223	174	184
Any Baseline Violent Arrest											
Has prior violent arrest	227	46%	2	8	14	9	123	182	224	175	184
No prior violent arrest	825	66%	4	10	14	10	126	176	223	177	187
Not Eligible for BAM - Upper Bound											
Not BAM Eligible	377	47%	2	8	14	8	118	166	208	162	171
BAM Eligible	675	71%	5	10	14	10	126	185	227	182	192
Not Eligible for BAM - Lower Bound											
Not BAM Eligible	221	44%	0	7	14	8	114	167	213	159	166
BAM Eligible	831	67%	5	10	14	10	127	182	224	180	190

Notes: This table produces the average number of CBT (SPARCS) sessions attended and the average number of YAP/mentoring received during the course of programming for participants in the program. We include the number of assigned treated youth by baseline characteristic, the percent who take-up by baseline characteristic, and the distribution of SPARCS and mentorship engagement among the participants.

Table III: C2C Arrest Outcomes, 6-12 Months post randomization

Outcome	Estimates				P-values		
	CM	ITT	CCM	TOT	Observed ITT	Observed TOT	FWER
Intensive Margin 6 Months							
Number of arrests	0.248	-0.013	0.170	-0.021	0.610	0.603	0.918
		(0.025)		(0.040)			
Number of violent arrests	0.064	-0.023	0.079	-0.037	0.017**	0.016**	0.078*
		(0.010)		(0.016)			
Number of property arrests	0.057	0.006	0.025	0.009	0.631	0.625	0.918
		(0.012)		(0.019)			
Number of drug arrests	0.017	0.013	0.000	0.022	0.076*	0.071*	0.271
		(0.008)		(0.012)			
Number of other arrests	0.111	-0.009	0.073	-0.014	0.566	0.559	0.918
		(0.015)		(0.025)			
Extensive Margin 6 Months							
Any arrest	0.161	-0.018	0.140	-0.029	0.199	0.192	0.484
		(0.014)		(0.022)			
Any violent arrest	0.060	-0.021	0.073	-0.035	0.017**	0.016**	0.084*
		(0.009)		(0.014)			
Any property arrest	0.049	0.000	0.028	0.000	0.987	0.987	0.988
		(0.009)		(0.014)			
Any drug arrest	0.012	0.012	0.000	0.019	0.032**	0.030**	0.122
		(0.005)		(0.009)			
Any other arrest	0.085	-0.008	0.065	-0.013	0.452	0.445	0.709
		(0.011)		(0.017)			
Intensive Margin 12 Months							
Number of arrests	0.490	-0.063	0.370	-0.103	0.115	0.109	0.397
		(0.040)		(0.064)			
Number of violent arrests	0.116	-0.026	0.115	-0.043	0.070*	0.066*	0.307
		(0.014)		(0.023)			
Number of property arrests	0.133	-0.019	0.099	-0.031	0.302	0.293	0.659
		(0.019)		(0.030)			
Number of drug arrests	0.035	0.002	0.016	0.003	0.861	0.858	0.860
		(0.010)		(0.016)			
Number of other arrests	0.205	-0.020	0.140	-0.032	0.384	0.376	0.659
		(0.023)		(0.036)			
Extensive Margin 12 Months							
Any arrest	0.244	-0.041	0.223	-0.067	0.008***	0.007***	0.041**
		(0.016)		(0.025)			
Any violent arrest	0.098	-0.025	0.098	-0.041	0.028**	0.026**	0.114
		(0.011)		(0.018)			
Any property arrest	0.095	-0.012	0.074	-0.020	0.291	0.283	0.505
		(0.012)		(0.019)			
Any drug arrest	0.025	0.002	0.009	0.003	0.755	0.751	0.762
		(0.007)		(0.010)			
Any other arrest	0.133	-0.018	0.113	-0.029	0.155	0.149	0.410
		(0.012)		(0.020)			

Notes: CM is the control mean. Intent-to-treat (ITT) and Treatment on the treated (TOT) estimates were calculated using randomization block fixed effects and robust standard errors. CCM is the control complier mean. We include the following baseline characteristics: demographic covariates (age/race/gender dummies), school grade at randomization indicators, prior enrollment in free/reduced lunch benefits, prior gap in enrollment indicator, prior arrest records by type (numbers and indicators), indicator for any prior victimization, and number of prior victimizations. FWER p-values in the last column. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table IV: C2C Arrest Outcomes, 24-36 Months post randomization

Outcome	Estimates				P-values		
	CM	ITT	CCM	TOT	Observed ITT	Observed TOT	FWER
Intensive Margin 24 Months							
Number of arrests	0.871	-0.054	0.635	-0.088	0.398	0.390	0.874
		(0.064)		(0.103)			
Number of violent arrests	0.226	-0.041	0.202	-0.066	0.076*	0.071*	0.331
		(0.023)		(0.037)			
Number of property arrests	0.218	-0.009	0.160	-0.015	0.730	0.726	0.983
		(0.027)		(0.043)			
Number of drug arrests	0.062	-0.001	0.034	-0.001	0.946	0.945	0.995
		(0.013)		(0.021)			
Number of other arrests	0.365	-0.003	0.240	-0.005	0.925	0.924	0.995
		(0.036)		(0.057)			
Extensive Margin 24 Months							
Any arrest	0.345	-0.063	0.328	-0.103	0.000***	0.000***	0.001***
		(0.017)		(0.027)			
Any violent arrest	0.164	-0.037	0.152	-0.060	0.009***	0.008***	0.030**
		(0.014)		(0.022)			
Any property arrest	0.137	-0.007	0.106	-0.011	0.634	0.628	0.865
		(0.014)		(0.022)			
Any drug arrest	0.044	0.003	0.020	0.005	0.717	0.712	0.865
		(0.008)		(0.014)			
Any other arrest	0.197	-0.017	0.164	-0.027	0.250	0.242	0.573
		(0.015)		(0.023)			
Intensive Margin 36 Months							
Number of arrests	1.175	-0.084	0.915	-0.136	0.285	0.277	0.818
		(0.078)		(0.125)			
Number of violent arrests	0.289	-0.029	0.245	-0.047	0.299	0.291	0.818
		(0.028)		(0.045)			
Number of property arrests	0.272	-0.009	0.210	-0.014	0.788	0.784	0.955
		(0.032)		(0.051)			
Number of drug arrests	0.090	-0.004	0.056	-0.006	0.812	0.809	0.955
		(0.016)		(0.026)			
Number of other arrests	0.524	-0.042	0.404	-0.069	0.327	0.319	0.818
		(0.043)		(0.069)			
Extensive Margin 36 Months							
Any arrest	0.386	-0.052	0.363	-0.084	0.004***	0.003***	0.018**
		(0.018)		(0.028)			
Any violent arrest	0.196	-0.031	0.179	-0.051	0.040**	0.037**	0.134
		(0.015)		(0.024)			
Any property arrest	0.159	-0.007	0.128	-0.011	0.641	0.636	0.872
		(0.015)		(0.024)			
Any drug arrest	0.059	0.004	0.032	0.006	0.694	0.689	0.872
		(0.010)		(0.015)			
Any other arrest	0.250	-0.033	0.229	-0.053	0.037**	0.034**	0.134
		(0.016)		(0.025)			

Notes: CM is the control mean. Intent-to-treat (ITT) and Treatment on the treated (TOT) estimates were calculated using randomization block fixed effects and robust standard errors. CCM is the control complier mean. We include the following baseline characteristics: demographic covariates (age/race/gender dummies), school grade at randomization indicators, prior enrollment in free/reduced lunch benefits, prior gap in enrollment indicator, prior arrest records by type (numbers and indicators), indicator for any prior victimization, and number of prior victimizations. FWER p-values in the last column. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table V: Discretionary Arrest Classification for All Arrests Post Randomization

FBI Category	N	Share of Arrests	% Non-Index	% Non-Index Excluding Simple A/B	% High Discretion using Majority Classification	% High Discretion using ML Model
Violent Arrests						
Non-negligent Homicide	22	0.67%	0%	0%	0%	0%
Negligent Homicide	1	0.03%	100%	100%	0%	0%
Sexual Assault	3	0.09%	0%	0%	0%	0%
Robbery	237	7.23%	0%	0%	0%	1.27%
Aggravated Assault	59	1.8%	0%	0%	0%	6.78%
Aggravated Battery	55	1.68%	0%	0%	0%	12.73%
Simple Assault	60	1.83%	100%	0%	0%	1.67%
Simple Battery	296	9.03%	100%	0%	0%	0%
Property Arrests						
Burglary	52	1.59%	0%	0%	0%	1.92%
Larceny	292	8.91%	0%	0%	0%	3.08%
Motor Vehicle Theft	332	10.13%	0%	0%	0%	29.52%
Drug Arrests						
Drug	240	7.32%	100%	100%	100%	80.83%
Other Arrests						
Forgery	2	0.06%	100%	100%	0%	0%
Fraud	4	0.12%	100%	100%	0%	50%
Vandalism	65	1.98%	100%	100%	0%	0%
Weapons Violation	463	14.13%	100%	100%	100%	68.03%
Prostitution	1	0.03%	100%	100%	100%	100%
Sexual Abuse	1	0.03%	100%	100%	0%	0%
Gambling	32	0.98%	100%	100%	100%	90.62%
Domestic Violence	7	0.21%	100%	100%	0%	42.86%
DUI	3	0.09%	100%	100%	100%	66.67%
Liquor License	2	0.06%	100%	100%	0%	50%
Disorderly Conduct	156	4.76%	100%	100%	0%	22.44%
All Other Offenses	329	10.04%	100%	100%	0%	20.97%
Chicago Municipal Code Violations	65	1.98%	100%	100%	100%	70.77%
Traffic Infractions	100	3.05%	100%	100%	100%	87%
Warrant Violations	398	12.15%	100%	100%	0%	43.22%
Total	3277		68%	57%	28%	32.93%

Notes: Table shows the results of our machine learning model on classifying high discretion arrests among all the outcome arrests for the study sample by FBI category (using the top charge of the arrest) and compares this to other existing methods of accounting for discretion. For example, we see that among the arrests where the top charge is a Chicago Municipal Code violation, 71% get classified as high discretion using our ML model (column 6). The first column shows the total number of arrests in that FBI categoryand the second column shows this number as a share of all outcome arrests. The third column presents discretion classification using the FBI non-index method. For example, all drug and charges under "other" arrests are considered non-index. The fourth column excludes simple assault and battery from the non-index/high discretion category. The fifth column uses our ML classification as a way to designate full categories of FBI charges as high-discretion if more than 50% of that charge in our ML model is predicted to be high discretion.

Table VI: C2C Discretionary Arrest Outcomes, Ever Arrested and Number of Arrests

	Index		Index Including A/B		Majority Classification		ML Model	
	Control Mean	Estimate	Control Mean	Estimate	Control Mean	Estimate	Control Mean	Estimate
Ever Arrested (High Discretion)								
6 months	0.112	-0.012	0.095	-0.009	0.044	0.003	0.046	0.001
		(0.012)		(0.011)		(0.009)		(0.009)
12 months	0.171	-0.024*	0.148	-0.023*	0.073	-0.004	0.084	-0.008
		(0.014)		(0.013)		(0.010)		(0.011)
24 months	0.255	-0.030*	0.217	-0.024	0.121	-0.010	0.133	-0.013
		(0.016)		(0.015)		(0.012)		(0.013)
30 months	0.286	-0.032*	0.245	-0.026*	0.147	-0.020	0.157	-0.017
		(0.016)		(0.015)		(0.013)		(0.013)
36 months	0.310	-0.035**	0.270	-0.036**	0.168	-0.025*	0.187	-0.028**
		(0.017)		(0.016)		(0.014)		(0.014)
Ever Arrested (Low Discretion)								
6 months	0.075	-0.005	0.096	-0.014	0.139	-0.028**	0.138	-0.020
		(0.011)		(0.012)		(0.013)		(0.013)
12 months	0.133	-0.024*	0.162	-0.032**	0.217	-0.048***	0.214	-0.045***
		(0.013)		(0.014)		(0.015)		(0.015)
24 months	0.193	-0.018	0.237	-0.034**	0.308	-0.053***	0.306	-0.055***
		(0.015)		(0.016)		(0.017)		(0.017)
30 months	0.208	-0.021	0.262	-0.039**	0.330	-0.052***	0.323	-0.050***
		(0.016)		(0.017)		(0.017)		(0.017)
36 months	0.225	-0.022	0.279	-0.033*	0.346	-0.041**	0.337	-0.038**
		(0.016)		(0.017)		(0.018)		(0.018)
Number of Arrests (High Discretion)								
6 months	0.156	-0.013	0.132	-0.007	0.057	0.006	0.065	0.004
		(0.020)		(0.018)		(0.012)		(0.013)
12 months	0.298	-0.041	0.250	-0.029	0.099	0.003	0.116	0.004
		(0.030)		(0.028)		(0.016)		(0.017)
24 months	0.541	-0.042	0.444	-0.019	0.191	-0.014	0.212	0.004
		(0.046)		(0.042)		(0.024)		(0.026)
30 months	0.660	-0.071	0.546	-0.051	0.241	-0.032	0.281	-0.025
		(0.052)		(0.047)		(0.026)		(0.030)
36 months	0.762	-0.077	0.636	-0.060	0.285	-0.032	0.340	-0.036
		(0.056)		(0.051)		(0.029)		(0.032)
Number of Arrests (Low Discretion)								
6 months	0.093	-0.006	0.116	-0.012	0.192	-0.025	0.183	-0.020
		(0.014)		(0.016)		(0.022)		(0.020)
12 months	0.194	-0.029	0.243	-0.041	0.393	-0.073**	0.374	-0.073**
		(0.022)		(0.025)		(0.035)		(0.033)
24 months	0.329	-0.016	0.426	-0.038	0.679	-0.043	0.659	-0.067
		(0.033)		(0.038)		(0.055)		(0.053)
30 months	0.378	-0.018	0.492	-0.039	0.797	-0.057	0.754	-0.070
		(0.037)		(0.043)		(0.063)		(0.059)
36 months	0.416	-0.014	0.542	-0.031	0.893	-0.059	0.836	-0.062
		(0.039)		(0.045)		(0.067)		(0.063)

Notes: These estimates show the impact of C2C on different methods of classifying high discretion arrests. The first set of results presents the RCT impacts using the discretion classification from the FBI index method. The second set of results excludes simple assault and battery from the non-index/high discretion category (includes simple assault and battery in the low discretion/index category). The third set of results uses our ML model as a way to designate full categories of FBI charges as high-discretion if more than 50% is predicted to be high discretion. The last set of results uses the individual arrest method of classifying high discretion from our ML model. ITT estimates were calculated using randomization strata, cohort level fixed effects, and robust standard errors. We include demographic covariates (age/race/gender dummies), school grade, prior enrollment in free/reduced lunch benefits, an ever not enrolled in school at baseline indicator, and prior arrest records (including prior high/low discretion arrest indicators). Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Figure I: C2C program effects on overall arrests, ITT

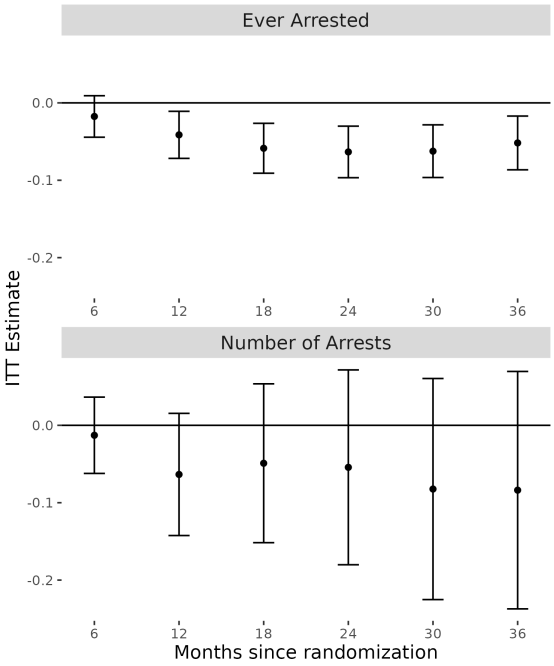


Figure II: C2C program effects on arrests for violent offenses, ITT

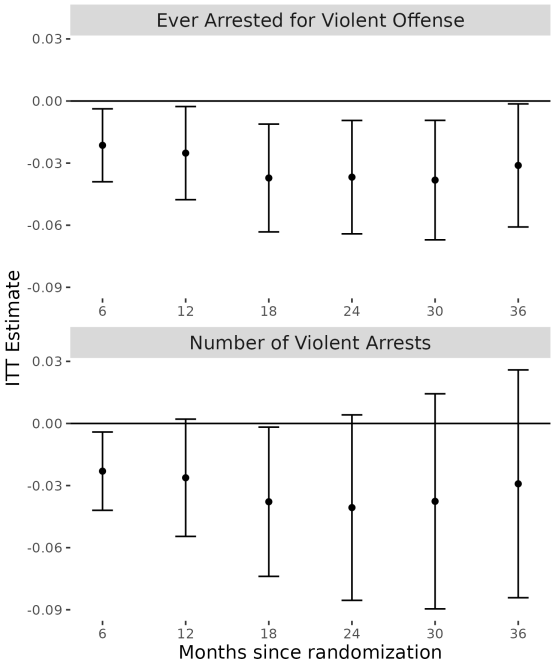


Figure III: Intensive and extensive margin effects on low discretion arrests, ITT

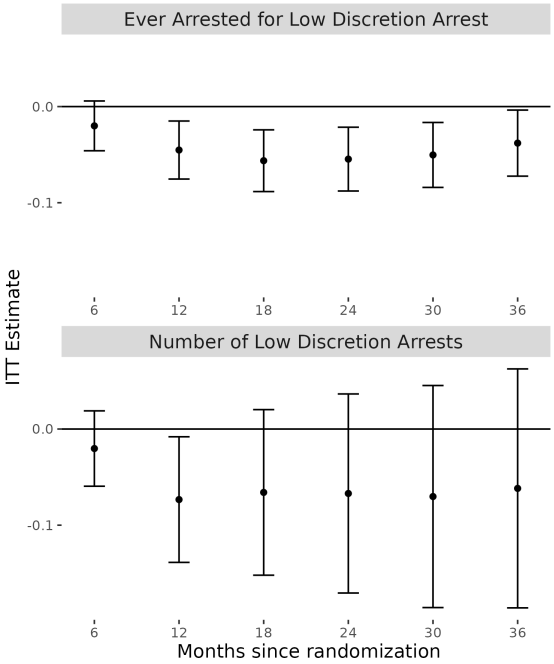
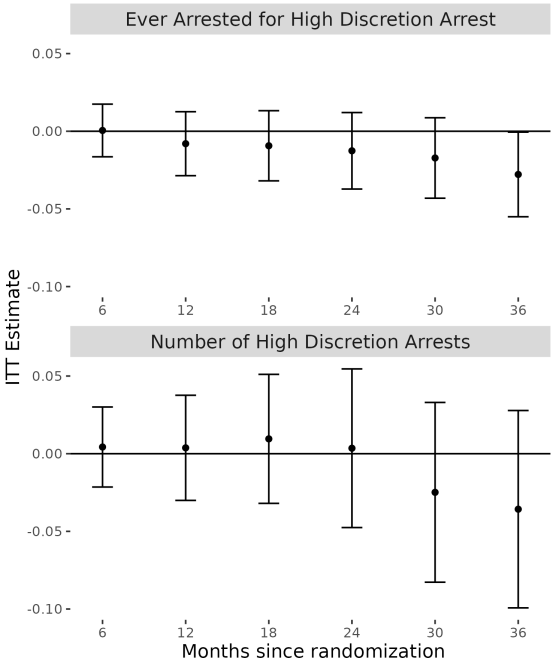


Figure IV: Intensive and extensive margin effects on high discretion arrests, ITT



Note: Figures shows the ITT estimates for different categories of arrests over time. The whisker lines represent 95% confidence intervals.

Appendix

A.1.1 Details on the C2C Program

Table A.I: C2C Program Elements, Wraparound Services

Wraparound Services	
Component	**Purpose**
Around-the-clock support and crisis intervention	YAP advocates are available 24/7 to provide immediate support when crises arise and help youth learn new ways of responding to challenges.
Creating a sense of security	The program helps youth and their family meet basic needs, as it includes consistent contact, care, wraparound services, and engagement with families to create a social safety net that youth can rely on even after their program participation ends.
New Experiences & group activities	New opportunities with peers can help youth process trauma by experiencing safety and relaxation in new environments. It also allows youth the opportunity to practice new ways of coping with peers that are working to make the same changes. C2C also exposes young people to new career opportunities and employment opportunities both inside and outside of their communities.
Individualized Support	YAP advocates formally identify young people's strengths, needs, preferences and goals across a spectrum of life domains to develop individualized service plans that become the basis of support.

Table A.II: C2C Program Elements, SPARCS†

SPARCS	
Topic	**Purpose**
Mindfulness exercises	Students are taught to be present, pay attention in a particular way on purpose and without judgement. They recognition and labeling of the link between emotion and the body, linking trauma to somatic symptoms. The goal is to affect regulation and impulsivity. An example of mindfulness exercises includes controlled breathing and weekly SOS (slow down, orient, self-check) practice.
Skill building: coping in the moment	Youth are given the tools to handle situations they can't immediately fix or change or in which action may worsen the situation. These tools involve distress tolerance skills such as distract and self-soothe techniques, identifying MUPS (coping strategies tat "Mess you UP") and defining ways these strategies may actually exacerbate or form new problems.
Skill building: Problem solving and creating meaning	Youth are taught skills to address past traumatic experiences or situations they do have control over. Youth are encouraged to construct a sense of purpose and meaning in their lives despite trauma. Utilizing LET'M GO (losing it, emotions, thoughts, meaning, goals and options) practice, they map out elements of reactions to an emotion situation and teach problem solving skills to address those emotions mindfully.
Skill building: Collaboration and communication	To address youth problems with alienation and trust youth learn and regularly practice communication skills through collaborative group work. The aim is also to assist youth in identifying and strengthening sources of social support. The MAKE A LINK technique provides a step-by-step guide so youth can more effectively manage their interpersonal interactions in any environment.
Immersive and experiential	SPARCS lessons are carried out in ways that tie in personal life experiences to better connect youth with the materials they learn about on paper. Lessons like, Portrait of my Life, helps participants learn about two key concepts, triggers and regulation of emotions, anger as well as the varying levels of intensity in which these feelings can occur.

†Structured Psychotherapy for Adolescents Responding to Chronic Stress

Figure A.I: Cohort 1-8 Communities Served

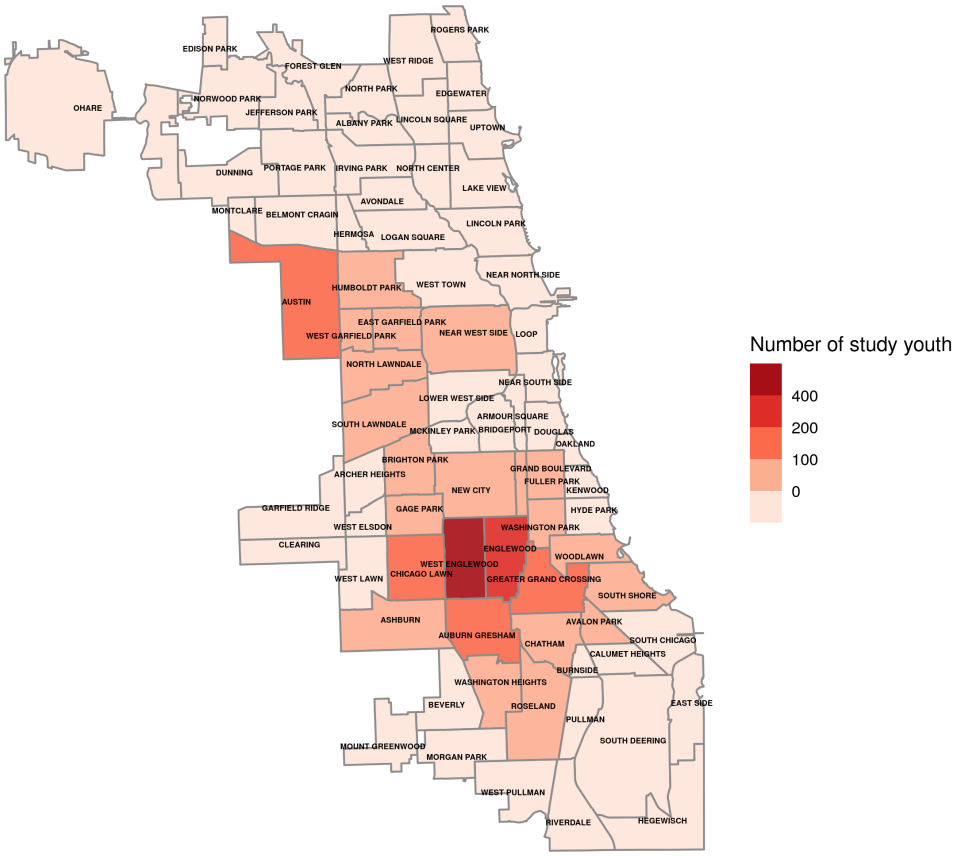
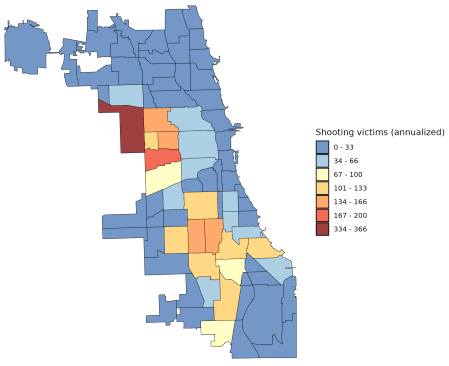
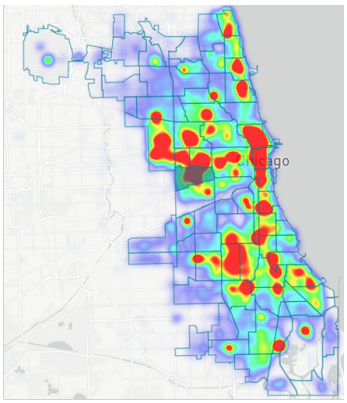


Figure A.II: Exposure to interpersonal violence and police



(a) Shootings by Community Area, Source: University of Chicago Crime Lab



(b) CPD Complaints, 32 year total, Source: Invisible Institute

A.1.2 Matching to CPD Data

To match our study participants to Chicago Police Department arrest data, we designed an algorithm using name, date of birth and address to match to Individual Record (IR) Numbers, a unique person-level identifier used by CPD . Specifically, if name and DOB are an exact match, we kept the match. However, given that CPD data is not accurate in every arrest record and that typos in name and DOB exist, we created several other pathways to ensure we would not miss accurate matches and designed the algorithm in a way to minimize false-positive arrests and manual review of particular cases. To generate potential matches, we match all CPD records to the C2C roster using DOB components (month/year, day/month, day/year) and compute a name similarity score for each match by calculating the Jaro-Winkler distance between each full name in the CPD data and C2C roster. Similarity scores are 0 if the names do not match at all, and 1 if they are an exact match. We consider any resulting match with a similarity score above .85 as a potential match. We allow for highly similar names (similarity score > 0.95) with a highly similar date of births (differing on regular time intervals that may be attributed to a typo such as one year/month/day) with a matching address. Matches were kept automatically if they have the same date of birth and same name or highly similar names and same date of birth. If the date of birth is the same, but the name match was not highly similar (similarity score < 0.95), the case was manually reviewed, taking into account address information. Matches that went to manual reviews were only kept if at least two independent reviewers (out of three) consider the match good. We considered these type of "low name matches" with exact date of birth because the algorithm only looks at string distance and this does not account for phonetically similar names that are spelled differently.

If the date of birth was highly similar, we went through a different set of decision nodes before any matches were kept. In this context, we accounted for name matches that were similar but also "common" names found in the CPD data base (specifically among the 100 most common first names or the 100 most common last names found in CPD arrest files for people born during or after 1995). If date of birth was highly similar yet the name is a common CPD name, the match went to manual review (or dropped if name was not highly similar by string distance function and the address does not match). If date of birth was highly similar but the name is not a common CPD name, then the match was kept if the name is highly similar and the address matches, or sent to manual review otherwise. Only low name matches and non-address matches were dropped. Lastly, matches where the date of birth was not highly similar were either dropped outright (given how closely the name matches) or sent to manual review based on address.

A.1.3 Transfers out of Chicago Public Schools, Impact of C2C on Incapacitation, and Data Censoring Overall

Table A.III: C2C Impact on Transfers out of Chicago and Incarceration

			ITT		TOT		
Outcome	N	Control Mean	Estimate	P-Value	CCM	Estimate	P-Value
Days Incarcerated							
6 months	2064	5.074	0.033	0.974	1.122	0.053	0.973
			(0.982)			(1.562)	
12 months	2064	11.130	-0.952	0.616	4.041	-1.541	0.610
			(1.897)			(3.017)	
18 months	2064	17.795	-1.793	0.541	7.675	-2.901	0.534
			(2.934)			(4.666)	
Days Transferred out of Chicago							
6 months	2064	6.572	-0.613	0.622	3.569	-0.993	0.615
			(1.242)			(1.975)	
12 months	2064	18.489	-1.253	0.657	13.201	-2.028	0.651
			(2.823)			(4.488)	
18 months	2064	32.169	-2.842	0.529	25.648	-4.600	0.522
			(4.517)			(7.181)	
Days Censored							
6 months	2064	11.645	-0.581	0.708	4.690	-0.940	0.703
			(1.548)			(2.461)	
12 months	2064	29.620	-2.205	0.508	17.242	-3.568	0.500
			(3.331)			(5.293)	
18 months	2064	49.964	-4.635	0.381	33.323	-7.501	0.373
			(5.293)			(8.411)	

Notes: Data is currently limited to 18 months post randomization. CM is the control mean. Intent-to-treat (ITT) and Treatment on the treated (TOT) estimates were calculated using randomization block fixed effects and robust standard errors. CCM is the control complier mean. We include the following baseline characteristics: demographic covariates (age/race/gender dummies), school grade at randomization indicators, prior enrollment in free/reduced lunch benefits, prior gap in enrollment indicator, prior arrest records by type (numbers and indicators), indicator for any prior victimization, and number of prior victimizations. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A.IV: C2C Victimization Outcomes, Ever Victimized and Number of Victimizations

		ITT		TOT		
Outcome	Control Mean	Estimate	P-Value	CCM	Estimate	P-Value
Any Violent Victimization						
6 months	0.072	-0.009	0.400	0.078	-0.015	0.393
		(0.011)			(0.017)	
12 months	0.125	-0.005	0.700	0.118	-0.009	0.695
		(0.014)			(0.023)	
24 months	0.206	-0.012	0.476	0.197	-0.020	0.468
		(0.017)			(0.027)	
36 months	0.289	-0.026	0.179	0.285	-0.042	0.172
		(0.019)			(0.031)	
Any Serious Violent Victimization						
6 months	0.036	-0.009	0.233	0.038	-0.015	0.225
		(0.008)			(0.012)	
12 months	0.060	0.002	0.876	0.042	0.003	0.874
		(0.010)			(0.016)	
24 months	0.115	-0.018	0.170	0.095	-0.029	0.162
		(0.013)			(0.021)	
36 months	0.169	-0.036	0.017**	0.155	-0.059	0.015**
		(0.015)			(0.024)	
Number of Violent Victimizations						
6 months	0.081	-0.010	0.416	0.085	-0.017	0.408
		(0.013)			(0.020)	
12 months	0.154	-0.013	0.471	0.148	-0.022	0.464
		(0.018)			(0.029)	
24 months	0.311	-0.047	0.133	0.299	-0.076	0.126
		(0.031)			(0.050)	
36 months	0.463	-0.068	0.088*	0.449	-0.111	0.083*
		(0.040)			(0.064)	
Number of Serious Violent Victimizations						
6 months	0.037	-0.006	0.460	0.036	-0.010	0.453
		(0.008)			(0.013)	
12 months	0.069	-0.001	0.907	0.055	-0.002	0.906
		(0.012)			(0.020)	
24 months	0.144	-0.031	0.086*	0.131	-0.051	0.080*
		(0.018)			(0.029)	
36 months	0.214	-0.054	0.014**	0.203	-0.088	0.012**
		(0.022)			(0.035)	

Notes: CM is the control mean. Intent-to-treat (ITT) and Treatment on the treated (TOT) estimates were calculated using randomization block fixed effects and robust standard errors. CCM is the control complier mean. We include the following baseline characteristics: demographic covariates (age/race/gender dummies), school grade at randomization indicators, prior enrollment in free/reduced lunch benefits, prior gap in enrollment indicator, prior arrest records by type (numbers and indicators), indicator for any prior victimization, and number of prior victimizations. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A.V: C2C Outcome Arrest Narratives, Number of Terms by Treatment Status

			ITT	
Number of Terms by Stage	N	Control Mean	Estimate	P-Value
Pre Cleaning	3090	175.9	-0.180	0.952
			(2.974)	
Stopwords + Signs Removed	3090	86.4	-0.580	0.689
			(1.449)	
Bigrams Added	3090	171.8	-1.160	0.689
			(2.899)	
Final Pruning	3090	86.3	-0.328	0.797
			(1.272)	

Note: N reflects total number of arrests (and arrest narratives) in the outcome period. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

A.1.4 Manual Classification of Arrest Narratives: High Discretion Arrests

With a sample of almost $N = 2,400$ outcome arrests among the C2C study population at the time of manual classification, we selected a random sample of $N = 600$ arrests to serve as our training data, and $N = 200$ arrests to serve as our test data for the machine learning model.[1a] The sample was stratified to maintain the same proportion of arrest types by charge level (violent, property, drug, or other) that we observe in the full sample. A team of four analysts manually classified the set of 800 arrests indicating which are high discretion arrests with each analyst classifying 200 narratives. All analysts were blind to the treatment status of the arrestee. After the first 20 narratives were classified, reviewers gathered and discussed the arrest manual classification protocol and specific scenarios to ensure consistent coding according to the classification protocol. After all narratives were classified, we duplicated this effort for a portion of the narratives with different reviewers to determine if we had consistent classification. We found a high degree of consistency in classification, with an average of 93 percent agreement between reviewers across the four different classifications.

[1a] A few arrests from this sample were dropped later due to an update in the way we matched our sample to CPD arrest data. Training data is the sample we use to tune our hyperparameters of the ML model. Testing data is used to test performance out of sample.

A.1.4.1 Protocol for Manual classification

We created an extensive protocol to support manual classification of the arrest narratives that were based on existing literature and Chicago context. Existing studies have shown that when the offense is serious and the available evidence is strong, police are more likely to arrest a youth (Myers, 2002; Sealock and Simpson, 1998). In contrast, almost 75 percent of arrests in our study sample are misdemeanors, and research has highlighted misdemeanor policing is often proactive, preventative and highly discretionary (Natapoff, 2017). Other research has focused on the extralegal factors and working rules police officers use to help them define suspicious people and places, which often rely on geography, race of suspect, time of day and age of suspect (Stroshine et al., 2008; Rosenfeld et al., 2012). For example, street stops in our sample were often justified if a young person was seen hanging out in a "hot spot" with known gang or drug activity. Research suggests that in areas with high crime and neighborhood disorder, police officers are more likely to anticipate danger and use authority or arrests (Skogan and Frydl, 2004). Research has also highlighted that police officers view policing as a "craft" and rely on the experience and skills acquired to determine how to respond to youth crime (Willis, 2013; Willis and Mastrofski, 2017). Recent work has shown how officer assignment (by neighborhood and seniority) mechanisms can impact police discretion, as Ba et al. (2021) found that in Chicago, non-index arrests (their definition of "discretionary" arrests) and use of force decline sharply with officer experience.

Proactive policing strategies, which have become the bedrock of many police departments, generate greater civilian-police contact by design, allowing for greater police discretion relative to standard reactive policing strategies (National Academies of Sciences et al., 2018). This discretion intersects directly with the racial bias present in policing (Goel et al., 2016; Goncalves and Mello, 2021; Weaver and Geller, 2019; Legewie and Fagan, 2019; Gelman et al., 2007). For example, Ba et al. (2021) have shown that white police officers in Chicago are 22 to 35 percent more likely than their non-white counterparts to make stops for vaguely defined "suspicious behavior," which disproportionately affects Black civilians. The Chicago Police Department has come under scrutiny in recent years for applying a range of controversial policing tactics, such as "Stop and Frisk", and use of force, including the killing of 17-year old Laquan McDonald (Force, 2016; of Justice, 2017). The 2017 Department of Justice (DOJ) report about CPD provides additional context on how hot spot policing and other policing strategies are used in Chicago. The DOJ report highlighted that CPD overrelies on specialized units for "hot spot" type policing, specifically that of specialized teams, such as the tactical (TACT), gang, saturation, and narcotics units. These units do not answer service calls, but aggressively seek out problematic activity by conducting traffic stops, making contacts, and effecting arrests (of Justice, 2017). One officer said "his TACT officers

like to hunt for offenders" and another remarked that "it's not called profiling, it's called being proactive." Another common strategy CPD is said to engage in is the policing tactic of "randomly stopping their police vehicle and opening one door; if anyone runs, an officer will get out and give chase; if no one runs, they will close the door and drive on." These tactics are confirmed in the existing research that highlights how police may use arrests to clear a corner, send a message in a high-crime neighborhood, or collect information (Stuart, 2016).

Lastly, minor traffic violations such as failure to signal a turn is often used as a pretext to run plates, check for warrant violations, or find evidence of other more serious wrongdoing. This is a police strategy that has often targeted minority communities for decades and is an established police strategy to control crime (Maclin and Saverese, 2018; Forman Jr, 2017). For this reason, we consider arrests that use pretext stops such as minor traffic violations as discretionary arrests.

Below we discuss several examples of arrest narratives found in our sample, and how our protocol classified them according to discretion.[2a] The first two incidences resulted in an arrest for a warrant violation. However, the first was classified as discretionary and the second was not. The first example details an arrest that was initiated when the civilian was pulled over for a minor traffic violation that is not specified in the narrative. The officer then suspects drugs are being used because of an odor coming from the car. They conduct a narcotic investigation and do not find any drugs but do find that a passenger of the vehicle had a warrant out for their arrest. Our protocol classified this event as a high discretionary contact arrest because the officer chose to initiate the contact that ultimately resulted in the arrest. In contrast, the second example details the arrest of a civilian in which the officers went to a specific residence after a call was made by the arrestee's mother because of an active warrant. The person was notified and subsequently arrested. This event was not classified as high discretion given our protocol.

1. *High Discretion.* "Above arrested in that he was the passenger of a gray dodge sedan bearing IL license plate#[plate num] which a/o's [arresting officer] observed commit a minor traffic violation while traveling n/b on the [number] block of [street]. A/o's then curbed said vehicle. As a/o's approached above vehicle, a/o's smelled a strong odor of cannabis emitting from the inside of the vehicle. A/o's, to conduct further narcotic investigation asked occupants out of vehicle. Leads name check of front passenger [arrestee's name] revealed there to be an active warrant under warrant#[warrant num]issued on [date]. [arrestee's name] was then placed into custody and transported

[2a]Some details are redacted to ensure confidentiality. Further more A/O means arresting officer, and R/O means responding officer.

to the [num] district for further processing. Warrant verified via leads desk [desk officer's name]#[num] at 2328hrs on [date] and a/o's were given hold#[num]. Further name check of [arrestee's name] revealed [arrestee's name] to have an active investigative alert under ia#[num]. Area north detective [detective's name] #[num]notified of active ia at 2334hrs on [date]. No further wants/warrants. No gipp/trapp. Isr completed"

2. *Low discretion.* "This is a fugitive apprehension unit arrest, in summary on the above date and time a/o's received a call from the mother of [arrestee's name]. After a/o's had made several attempts to make contact with [arrestee's name] on said warrant. The mother stated that her daughter was wanted on warrant [number]. At which time a/o's went to the listed address and observed [arrestee's name]. She was placed into custody on said warrant and transported to the [district number] dist for further processing. Shows no parole status. Not assigned to g.i.p. denies any gang affiliation, no investigative alerts."

3. *High Discretion.*"In summary; a/o's on aggressive patrol assigned to violence zone [x] due to the recent gang related shootings in said area. A/o's observed the offender loitering on the sidewalk of [address] a hangout for the [faction name] faction of the [gang name]. A/o's conducted an investigatory stop and gang dispersal of the offender and co-arrestees under event #at [x]hrs. The offender was ordered to disperse and not to return from within sight or hearing for the next 8 hours. A/o's returned to the same area and observed the offender loitering in the same location at [address] with the co-arrestees whom are all self admitted and documented [gang name]. The offender was then placed in custody and transported into the 00xth district for further processing. Name check and investigative alert cleared. Offender is a juvenile and his guardian/aunt was notified at X hrs via phone."

4. *Low Discretion.* "Event #xxxxx this is a bwc [body worn camera] arrest. In summary, a/o's responded to a person shot call at the address of [address] recorded under rd#xxxxx. A/o's, while walking up to residence from approximately [address], observed above offender [offender's name], who is previously known to r/o's to be a member of the [gang name] street gang and from being a gun shot victim, exiting the front porch of the residence and walking down the front porch. A/o's observed offender [offender's name] to be wearing a black t-shirt and red pulled up jogging pants with a large bulge in the front waistband area consistent with a firearm. A/o's approached offender for a field interview. A/o [officer] performed a protective pat down of offender and felt a hard metal object consistent with being a firearm. Above detained. R/o

[officer] then recovered from offenders front waistband (1) glock 22, .40 caliber handgun, serial #xxxx with a 4 inch barrel, loaded with 12 live rounds and one live round in the chamber(inv#xxxxx). Offender placed in custody, mirandized, and transported into 00Xth district for processing. Grandmother, [grandmother's name] [number] who is offenders legal guardian, notified on scene. Has no i.d. Name check clear. No warrants/alerts on file. Not on parole.."

The detailed protocol we followed (with more examples) for our manual classification can be requested.[3a]

A.1.5 ML Discretionary Arrest Model

A.1.5.1 Overview of ML Model

We utilize a random forest model for our supervised machine learning classification process for high discretionary arrests. Random forest models are a tree-based classification method that avoid overfitting by averaging predictions from many trees that have been grown from a random subset of predictors. Tree-based classification methods seek to predict y from a feature vector x that divide the feature space into rectangular regions, and then fit a simple model in each rectangle.[4a] An individual tree will not be particularly good at prediction and suffer from high variance (a small change in the underlying data can lead to different sequence of splits and hence a different prediction), but the main intuition behind random forest (and other decision forests) is to build many trees and classify by merging their results. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Decision trees have become a popular nonlinear approach for text analysis as they are flexible, capture nonlinearity dependencies and incorporates rich interactions into classification in a simple and interpretable way (Breiman et al., 1984; Gentzkow et al., 2019). Random forests are also recommended in scenarios where you have sparsity: settings with a large number of features that are not related to the outcome. This is especially relevant in our situation given the large number of text features we have from the arrest narratives. Random forests are effective at picking up on the sparsity and ignoring the irrelevant features even if there are many of them (Athey and Imbens, 2019). Lastly, random

[3a] Due to space limitations, it was not included in this Appendix.

[4a] For a more intuitive analogy think about a real-life decision tree that models a set of sequential and hierarchical decisions for some final result; see Mullainathan and Spiess (2017) for an overview of common ML models used in economics and the intuition behind the models.

forests also require relatively little tuning of model parameters and have great performance "out of the box."[5a]

The main predictors of the model we use are text features (unigrams and bigrams) extracted from the arrest narratives, but we also included a standardized subset of non-text covariates: race, gender, age at the time of arrest; top charge, top charge's FBI code, second charge (if any), third charge (if any), number of additional charges; arrest beat; arrest year, arrest month; and indicators for an arrest being labeled as "other", "violent", or "drug" related.

To build our ML model, we first do a regularization step in which we select the optimal subset of text features to include in the model. We then select the optimal number of predictors to be randomly sampled in each tree (the mtry hyperparameter). Lastly, we select the optimal classification decision threshold.[6a] We describe each step in detail below.

A.1.5.2 ML Details: Model Selection and Tuning Parameters

To build our model we first do a regularization step, in which we select the optimal subset of text features in a stepwise way, adding text features in a sequential order using 4 different criteria.[7a] We then tune the mtry hyperparameter, selecting the optimal number of predictors to be randomly sampled with each tree. We describe this process in detail below:

1. First, given the high dimensionality of the text data from the arrest narratives, we reduced the number of text features to something that is more manageable. We employed a common standard text mining process that first removes punctuations, single numbers, common English stop words (e.g., "and", "or", "me", "when", etc.), and then uses word stemming (for example, "runs" and "running" are both replaced with the stem word "run") to standardize the remaining features.[8a] The next step in our process was to represent the narratives using what is called a *bag-of-words*. Here, the order of words was ignored and we created a matrix, c_{ij}, with dimensions determined by the number of surviving text features in the entire universe of narratives, and the number of distinct narratives included in the sample. The dimensionality of c_{ij} can grow exponentially with the number of narratives included, given that each narrative is usually a large and complex entity by itself. Matrix dimensionality can also grow

[5a] We initially explored other models but quickly landed on random forest as the best possible model for our purposes.

[6a] For the other parameters in the random forest model such as node size and max nodes we use the default numbers provided in the R randomForest package, package manual can be found `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`.

[7a] All non-text features are always included in the model.

[8a] We use the quanteda package in R.

when we allow for multiple words to be combined together into a single feature; this is why we limited the number of words to be bundled together to a maximum of two - i.e. we build our bag-of-words using only unigrams and bigrams. Lastly, we kept the 1,000 most frequent terms, provided that they appear on at least five arrest narratives in our sample. Before we started the formal regularization stage with the surviving 1,000 text features, we created an additional measure to help us filter more text. We created what is knows as the "term frequency–inverse document frequency" weights (tf-idf), an approach that can help exclude both common and rare words.[9a]

2. We then evaluated four different methods or family of models aiming to find the one that seems to better predict high discretionary arrests. Within each of these methods, we tuned the mtry parameter. First we chose a fixed mtry parameter, and varied the text features.[10a] The performance metric we used to evaluate model performance on each iteration is the area under Curve (AUC) from the receiver operating characteristics (ROC) curve. A ROC curve plots the true positive rate against the false positive rate for each model, varying the decision thresholds. The ROC curve is a frequently used tool for simultaneously showing both types of errors for all possible decision thresholds and can capture the predictive potential of each model. Binary classifiers are typically evaluated with ROC curves given you can depict the relative trade-offs that a classifier makes between benefits (true positives) and costs (false positives) (James et al., 2013). An ideal ROC curve will hug the top left corner (high true positive rate and a low false positive rate), so the larger the AUC the better the classifier.[11a]. The four types of models considered were:

   (a) Variable importance: after the standardized text mining process described above, text features are ranked by their estimated variable importance (using a mean decreased accuracy criteria). This rank is then used to add terms sequentially into the model as predictors, evaluating performance with each new addition in a standardized way. We then use this performance metric to select an optimal subset of text features to include (i.e. how many and which ones). A variant of this model converts all tf-idf weights into dummy variables. Note that we can only

[9a] Specifically, for a word j in document i, term frequency (tf_{ij}) is the count c_{ij} of occurrences of j in i. Inverse document frequency (idf_j) is the log of one over the share of documents containing j. The measure which is the product, tf_{ij} x idf_j, will have low scores for rare words because tf_{ij} will be low, and very common words that in most or all documents will have a low score because idf_j will be low.

[10a] The fixed mtry parameter was defined as: $m = 2 * sqrt(N)$, where N is the the number of predictive features (text + non-text) included in the model on each iteration.

[11a] An AUC of roughly 0.5 means that the classifier performs no better than chance (the model has no skill). AUC's of 0.85-0.90 and above are considered very good models (James et al., 2013; D'Agostino et al., 2018)

estimate variable importance using narratives that have already been classified. For the present case we do that using the train set, only.

(b) TF-IDF: after the standardized text mining process described above, text features are ranked by their average tf-idf weights, across all narratives. This rank is then used to add terms sequentially into the model as predictors, evaluating performance with each new addition in a standardized way. We then use this performance metric to select an optimal subset of text features to include (i.e. how many and which ones) . A variant of this model converts all tf-idf weights into dummy variables.

(c) Chi-squared: after the standardized text mining process described above, text features are ranked by the chi-squared statistic obtained in between a given feature and the outcome variable (i.e. the discretionary arrest indicator in this case). This rank is then used to add terms sequentially into the model as predictors, evaluating performance with each new addition in a standardized way. We then use this performance metric to select an optimal subset of text features to include (i.e. how many and which ones). A variant of this model converts all tf-idf weights into dummy variables.

(d) Term frequency: after the standardized text mining process described above, text features are ranked by their overall term frequency, across all narratives. This rank is then used to add terms sequentially into the model as predictors, evaluating performance with each new addition in a standardized way. We then use this performance metric to select an optimal subset of text features to include (i.e. how many and which ones). A variant of this model converts all tf-idf weights into dummy variables.

3. Once the optimal subset of text features has been determined, we then varied the mtry parameter, using values ranging from 1 to 500 (or until each model's max number of features as reached). We used this approach instead of simultaneously varying both text features and mtry parameter and optimizing over both due to the unreasonable amount of computing time it would take to run all the different possible model simulations. We recognize we may not be choosing the optimal text and mtry combination, but given the performance of the models we test we are satisfied with this approach.

4. After completing the regularization and parameter tuning steps for each model, we were interested in getting a sense of the AUC's distribution. In order to estimate the distribution in the train set, we used a single out-of-the-bag prediction using our

regularized/tuned models, and then bootstrapped that prediction 100,000 times estimating the AUC on each iteration. The mean and confidence intervals coming from the bootstrap process are presented on Table A.VI.

5. Finally, in order to estimate the AUC distribution in the test set, we also used single prediction coming from our regularized/tuned models, but this time we estimated the model using our train set data only, and then made an out-of-sample prediction in the test set. Using that prediction, we followed the exact same bootstrap procedure as we did with the train set. These results are also shown on Table A.VI.

Table A.VI below shows the AUC distributions in the train and test set using optimal parameters for each method (i.e. a regularized subset of text features + a tuned mtry parameter).

Table A.VI: Summary Table - Model Performance Results using an AUC criteria

	MODEL PARAMETERS			TRAIN SET (N=591)		TEST SET (N=196)		TRAIN SET (N=591)	TEST SET (N=196)
Model	**tf-idf as dummies?**	**optimal number of text features**	**optimal mtry**	**mean(AUC)**	**AUC's 95% conf. interval**	**mean(AUC)**	**AUC's 95 % conf. interval**	**Rank**	**Rank**
Var Imp. model 2	yes	388	13	0.9357	[0.9149 , 0.9536]	0.9258	[0.8801 , 0.9612]	1	3
Var Imp. model 1	no	414	10	0.934	[0.9131 , 0.9519]	0.9273	[0.8822 , 0.9625]	2	1
TF-IDF model 2	yes	873	47	0.9262	[0.9031 , 0.9462]	0.9202	[0.8732 , 0.9567]	3	4
TF-IDF model 1	no	875	50	0.9252	[0.9017 , 0.9455]	0.9189	[0.8718 , 0.9558]	4	5
Chi2 model 1	no	307	9	0.9242	[0.9009 , 0.9444]	0.9261	[0.8809 , 0.9611]	5	2
Term Freq. model 2	yes	911	76	0.9146	[0.8884 , 0.9372]	0.9175	[0.8708 , 0.9538]	6	8
Term Freq. model 1	no	803	70	0.9077	[0.8814 , 0.9308]	0.9181	[0.8715 , 0.9545]	7	7
Chi2 model 2	yes	422	33	0.9065	[0.8795 , 0.9305]	0.9186	[0.8718 , 0.9550]	8	6

Table A.VII: Summary Table - Model Performance Results using a precision/recall ROC

	MODEL PARAMETERS			TRAIN SET (N=591)		TEST SET (N=196)		TRAIN SET (N=591)	TEST SET (N=196)
Model	**tf-idf as dummies?**	**optimal number of text features**	**optimal mtry**	**mean(AUC)**	**AUC's 95% conf. interval**	**mean(AUC)**	**AUC's 95 % conf. interval**	**Rank**	**Rank**
Var. Imp. model 2	yes	388	13	0.8443	[0.7923 , 0.8893]	0.7384	[0.5742 , 0.8782]	1	2
Var. Imp. model 1	no	414	10	0.8367	[0.7832 , 0.8831]	0.7331	[0.5667 , 0.8818]	2	3
Chi2 model 2	yes	422	33	0.8345	[0.7801 , 0.8815]	0.7324	[0.5727 , 0.8627]	3	4
Chi2 model 1	no	307	9	0.8318	[0.7759 , 0.8803]	0.7302	[0.5650 , 0.8777]	4	7
TF-IDF model 2	yes	873	47	0.8278	[0.7714 , 0.8766]	0.7322	[0.5689 , 0.8685]	5	5
TF-IDF model 1	no	875	50	0.8239	[0.7666 , 0.8738]	0.7246	[0.5603 , 0.8629]	6	8
Term Freq. model 1	no	803	70	0.8235	[0.7662 , 0.8730]	0.7315	[0.5752 , 0.8582]	7	6
Term Freq. model 2	yes	911	76	0.8222	[0.7650 , 0.8716]	0.7402	[0.5892 , 0.8609]	8	1

Table A.VII was constructed in the same way as Table A.VI, with the sole exception that all AUCs shown now come from precision-recall ROC curves. Precision (also known as positive predictive value) is a ratio of the number of true positives divided by the sum of the true positives and false positives. Recall is the true positive rate. Precision-recall

curves describes how good a model is at predicting the positive/minority class (in this case discretionary arrests). The AUC under these curves varies between 1 (best performance) and 0 (poor performance); a precision value of 1 for a given classifier would mean that all real positive cases were predicted by the model, with no false positive cases altogether. Precision-recall ROC curves can be a helpful tool in evaluating the performance of a model when there is imbalance in the observations between the two classes (for example, when there are few examples of a class), or when we are less interested in the skill of the model in predicting the majority class, in this case non-discretionary (e.g. high true negatives) (Saito and Rehmsmeier, 2015).[12a] We used the PR-AUC as a helpful metric when evaluating the already regularized and tuned models.

A.1.5.3 Optimal Model

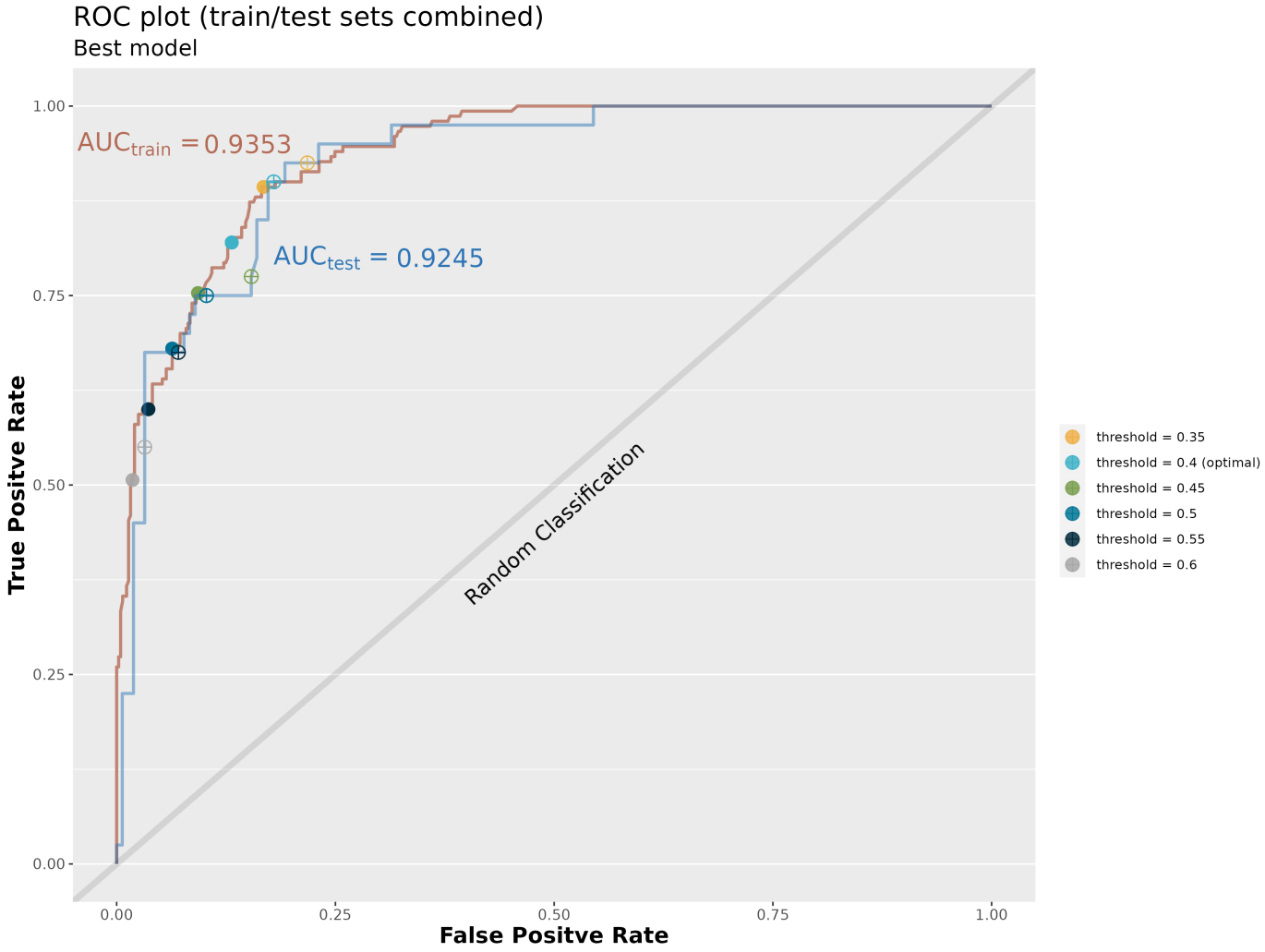
The classification model that shows the better overall performance (using both performance metric curves) in the train & test sets is the **variable importance model 2 (with dummies)**. Based on these results, we use this model to make a final prediction for the discretionary arrest indicator in the unclassified sample.

A.1.5.4 Optimal Model's TPR/FPR ROC

Figure A.III below shows the true positive-false positive rate ROC curves obtained for the optimal model, both when predicting within the train set (red line), and test set (blue line). For each case, we've included a number of different classification thresholds, signaled by the colored dots along the respective ROC curves. These dots show the estimated true positive/false positive coordinates, when varying the acceptance threshold to classify a given narrative as discretionary = high, based on the model's predicted class probabilities. The grey line in the plot depicts a ROC obtained when using a random classifier; note that this ROC resembles a 45 degree line, with an AUC of 0.5.

[12a] Precision-recall AUCs shown in Table A.VII are systematically lower than the AUCs shown in Table A.VI. This can be explained by the fact that a random classifier in a precision-recall curve would produce an AUC that's similar to the percentage of real positive cases found in the data. In the case of the discretionary arrest indicator, that percentage is 24.14% in the manually classified sample we use. That is, the expected AUC when using a random classifier in the precision-recall cases is ∼0.2414, significantly smaller than the expected AUC of 0.5 for a random classifier in the true positive-false positive comparison/ROC curve. Because a random (sometimes call unskilled) classifier is an important benchmark when evaluating how good a model is performing, it might be the case that the precision-recall AUCs shown in Table A.VII provide an even better improvement, if compared to a prediction at random, than those shown in Table A.VI.

Figure A.III

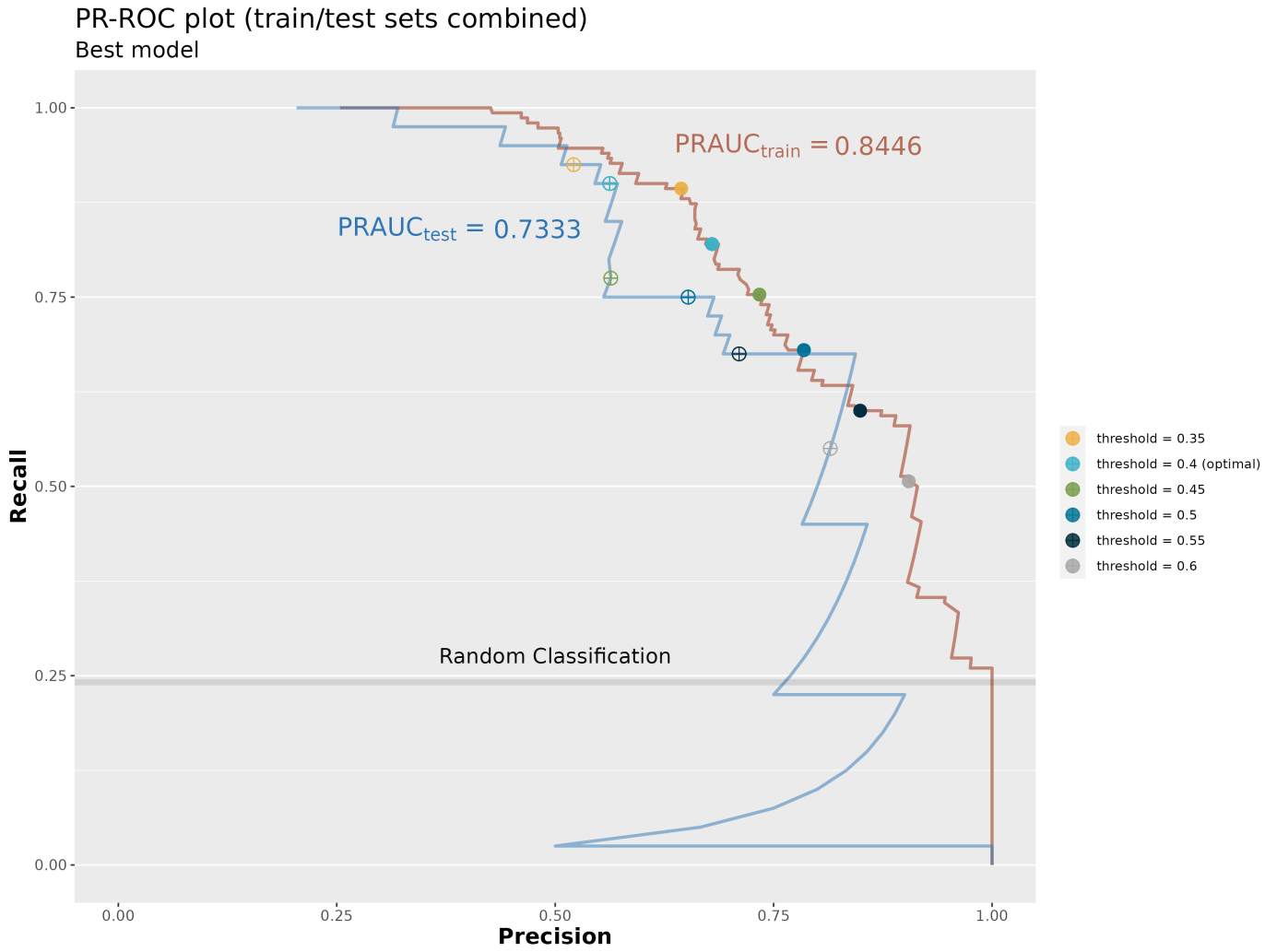


A.1.5.5 Optimal Model's precision/recall ROC

Figure A.IV shows the precision-recall ROC curves obtained for the optimal model, both when predicting within the train set (red line), and test set (blue line). For each case, we've included a number of different classification thresholds, signaled by the colored dots along the respective ROC curves. These dots show the estimated precision/recall coordinates, when varying the acceptance threshold to classify a given narrative as discretionary = high, based on the model's predicted class probabilities. The grey line in the plot depicts a precision-recall ROC obtained when using a random classifier; note that this PR-ROC curve

resembles a horizontal line that intersects the y-axis at the fraction of positive cases in the sample (0.2414). [13a]

Figure A.IV



A.1.5.6 Decision threshold selection

The last key decision is selecting the optimal decision threshold cutoff. This parameter, which we tuned using a bootstrap methodology, uses the predicted probability or scoring of

[13a] This means that the random classifier's AUC is lower in this case, if compared to the true positive-false positive rate case (~0.25 vs 0.5).

class membership and converts it to a class label (i.e. high or low discretion).[14a] We use various metrics to help us choose the optimal decision threshold using the selected model: G-mean, Youden's J, and F-measure. For each point in a standard ROC curve, the G-mean is defined as: $G_i = \sqrt{tpr_i} \cdot (1 - fpr_i)$, by optimizing this metric we find the point on the ROC curve that seeks a balance between the true positive rate and the false positive rate. An alternative metric is the J statistic which is calculated as $J_i = tpr_i - fpr_i$. And lastly, we also look at a metric with the precision-recall curves that seeks to balance precision and recall, the F-measure, which is calculated as $F_i = 2\frac{prec_i \cdot rec_i}{prec_i + rec_i}$.[15a] We decided to move forward with a **0.4 classification threshold** when making predictions in the unclassified sample as this threshold resulted in the most optimal model performance when looking at the test set results. This essentially means we'll use the estimated model probabilities to classify each narrative as discretionary = high/low in the following way: classify as high if Prob(disc = high) ≥ 0.4.

A.1.5.7 Optimal model's prediction on unclassified sample

Table A.VIII below shows the distribution of discretion = high when using our optimal model to predict in the unclassified study sample of C2C outcome arrests (N=3,277), and how that compares to the observed distribution of the same class in the test/train sets (i.e. manually classified). We also show the results when predicting in the C2C baseline arrest sample (N=3,087). Note that the share of arrests predicted as discretion = high is larger in the unclassified outcome arrest sample (32.93%) vs the manually classified one (∼24% when we bundle the train/test sets together). We think this could be explained by a few different factors. First, the train/test set data under-represents arrests in the other/drug categories,[16a] and the incidence of discretionary arrests is significantly larger in those arrests vs. arrests labeled as violent/property (based on manually classified narratives we see and incidence of ∼8.5% in the latter vs. ∼40% in the former). Second, the differences could be also due to statistical sample variations - i.e. it could be the case that the unobserved incidence of discretionary arrests in the unclassified sample is indeed a bit higher than the

[14a] For example, a 50% classification rule (the default in most models) means that we'd only classify a given narrative as discretionary = "high" if the predicted class probability is greater than or equal to 0.50.

[15a] A list of optimal thresholds based on the first two metrics were obtained after bootstrapping 100,000 standard ROCs using a single prediction from our optimal model; on each ROC, we maximized the given criteria and found the corresponding optimal classification threshold. Another list of optimal thresholds was produced based on the third metric (i.e. the F-measure), again obtained after bootstrapping 100,000 PR-ROCs from a single prediction coming from our optimal model. In the same way, on each PR-ROC we maximized the F-measure in order to find the corresponding optimal threshold. We do this using predictions obtained both in the train set and the test set.

[16a] Indeed, the train/test set data uses a 50/50 split sample of other/drug arrests vs. violent/property arrests; but in the full arrest outcome sample the split is closer to 52/48

one we observe in the train/test set. Finally, the fact that we're using a "more permissive" classification threshold - i.e. more permissive than a majority vote rule would be - could also be factoring into the discretion = high rate we observe in the predicted outcome arrests sample: at the end, the metrics used to select an optimal threshold balance the trade-offs in between true/false positive rates.

Table A.VIII: Summary Table - Prediction in the unclassified sample (baseline + outcome arrests)

Sample	**disc = high***	**Obs**	**Classification Method**
Train Set (outcomes only)	25.38%	591	Manual
Test Set (outcomes only)	20.41%	196	Manual
Other outcome narratives	35.7%	2490	Predicted
All outcome narratives	32.93%	3277	Mixed
All baseline narratives	17.65%	3087	Predicted
Full arrest sample	25.52%	6364	Mixed

*Predicted classification was done using a tuned decision threshold criteria. We find that the optimal classification threshold corresponds to classifying as disc ="high" if the predicted probability for class "high" in a given narrative is ≥ 0.4

A.1.5.8 Performance Measures and Robustness Checks

The first two rows of table A.IX display model performance measures on the selected optimal model for predicting discretionary arrests on the train and test set. We see from the out-of-the-bag (OOB) estimates (predictions obtained by averaging tree-specific predictions on a subset of observations that were deliberately left out when growing them) from the train set that the optimal model was able to achieve a high AUC ($\sim$0.94). We see this manifested in a high true positive rate (82%). Our test set shows relatively similar performance with an AUC of $\sim$0.92 and a true positive rate of 90%. [17a]

Table A.IX: Classification Performance Metrics - Discretionary Arrests

Subset	Mean AUC	Miss-classification Rate	True Positive Rate	False Positive Rate	Log-loss	RMSE	Obs
Train Set	0.9353	0.1438	0.8200	0.1315	0.3376	0.3792	591
Test Set	0.9245	0.1633	0.9000	0.1795	0.3455	0.4041	196
Random CPD Sample	0.8885	0.2000	0.8800	0.2267	0.3789	0.4472	100
Random CPD Sample similar to C2C	0.8625	0.2000	0.9286	0.2931	0.4446	0.4472	100

To assess the robustness of our optimal model, we constructed two random samples of arrest narratives using arrests that are not in our C2C evaluation sample. These samples

[17a] Although we followed best practices in the literature and optimized on the AUC, we show the other performance measures for transparency and for ease of interpretation.

were defined as: (i) a random subset of 100 CPD arrests that happened in 2015 or after; (ii) a random sample of 100 CPD arrests that happened in 2015 or after, stratified to be representative of our C2C evaluation sample in a number of relevant dimensions such as age at the time of arrest, race, gender, and arrest location.[18a] We then manually classified all narratives within both samples as discretionary = high/low, using the same classification criteria we had applied to the C2C evaluation sample. The final step on this robustness check process included testing our optimal model by making predictions on both random samples mentioned above. We find similar true positive rates and misclassification rates for these two random subsets. The AUC is slightly lower than the test/train set samples (AUC around 0.86 and 0.89 for each of the samples, with the random CPD sample doing slightly better), however the model is still considered to be well performing in these samples.[19a] Table A.IX also displays alternative measures of model performance. Based on existing literature (Bradley, 1997; Hand, 2009; Saito and Rehmsmeier, 2015) and given the classification problem we were working with, we decided to use the area under the ROC and PR-ROC curves as primary performance metrics with our models.

These robustness checks (along with the test set results) confirmed we did not overfit our ML model in our study data. It also demonstrates the ability to classify these behaviors in completely new arrests samples. We believe this robustness check also helps demonstrate the usefulness of this new approach and suggests its potential usefulness in other applications.

As a robustness check, we wanted to confirm that if we vary the decision threshold, the results remain qualitatively very similar (with no movement on discretionary arrests, and a reduction on non-discretionary arrests). As discussed earlier, what we found through our analysis is that a 40% threshold is optimal in that it balances between achieving a high true positive rate and low enough false positive rate. However, we'd like to confirm that our findings are not sensitive to this threshold cutoff had we had less (or more) tolerance for false positives or true positives, within a reasonable range. In Figure A.V and A.VI we confirm this. We vary the decision threshold of the model (from 0.30 to 0.60 at 0.02 increments), re-run our predictions for the arrest sample for each new threshold, and then look at extensive effects at 12 and 24 months post randomization. We find that the low discretionary arrest estimate is very stable and highly significant regardless of the threshold cutoff. The high discretionary arrest estimates are sometimes a bit noisier, but in general close to zero with large confidence intervals. We do generally see the high discretionary estimate grow smaller as the threshold increases because. Likewise, as would be expected, as we increase the threshold, we predict a smaller portion of the arrest sample as high

[18a] From here onward we'll refer to this sample as the "similar to C2C" random sample.

[19a] AUCs in the 0.8-0.9 range would correspond to good/moderate classification accuracy.

discretionary. The 50% threshold predicts about 18% of the sample as discretionary (in comparison to what we use in main estimates which is 25.5%). Our goal was to maximize AUC, not minimize the misclassification rate. However, we show in Figure A.VII had we chosen the decision threshold based on the lowest misclassification rate (0.57) our findings would hold as well. The consistency in our findings highlights again that our main RCT findings are being driven by reductions in low discretionary arrests, with little discernible impact on high discretionary arrests.

Figure A.V: C2C program effects on any arrest 12 months post-randomization, ITT

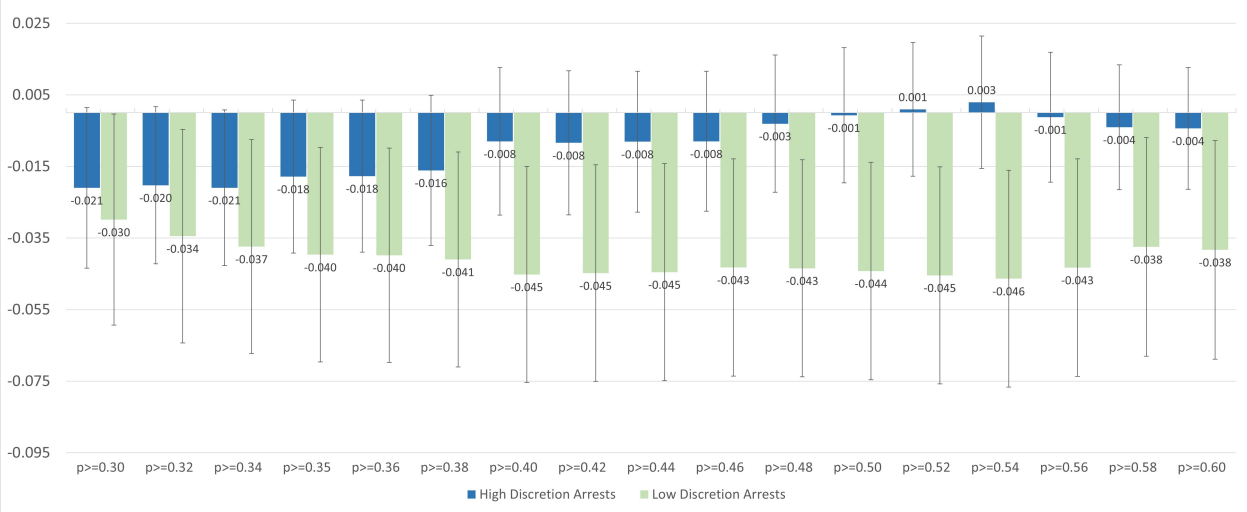


Figure shows the ITT estimate as a function of the classification threshold and the 95% confidence interval

Figure A.VI: C2C program effects on any arrest 24 months post-randomization, ITT

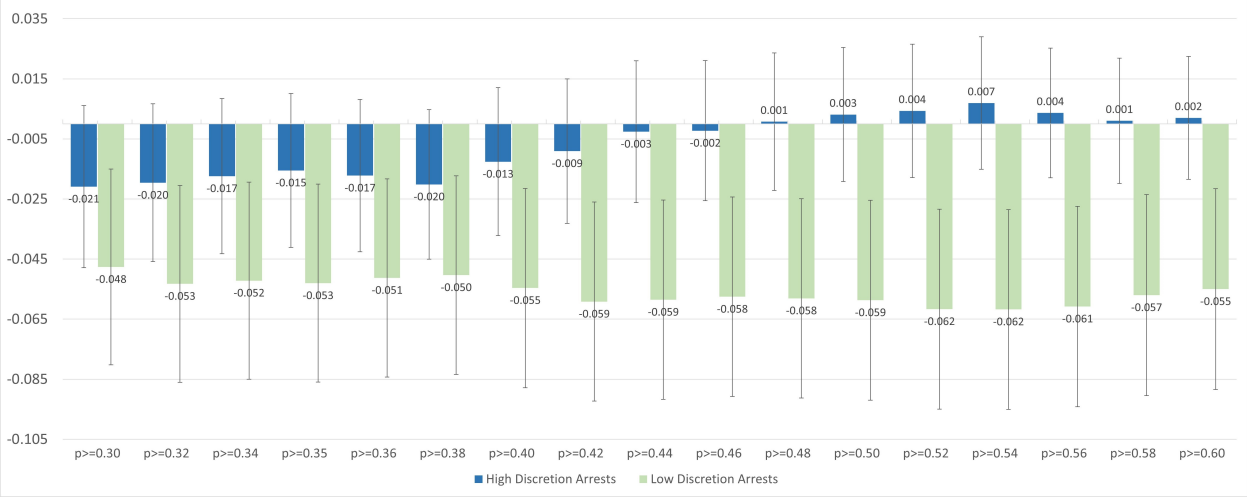


Figure shows the ITT estimate as a function of the classification threshold and the 95% confidence interval

Figure A.VII

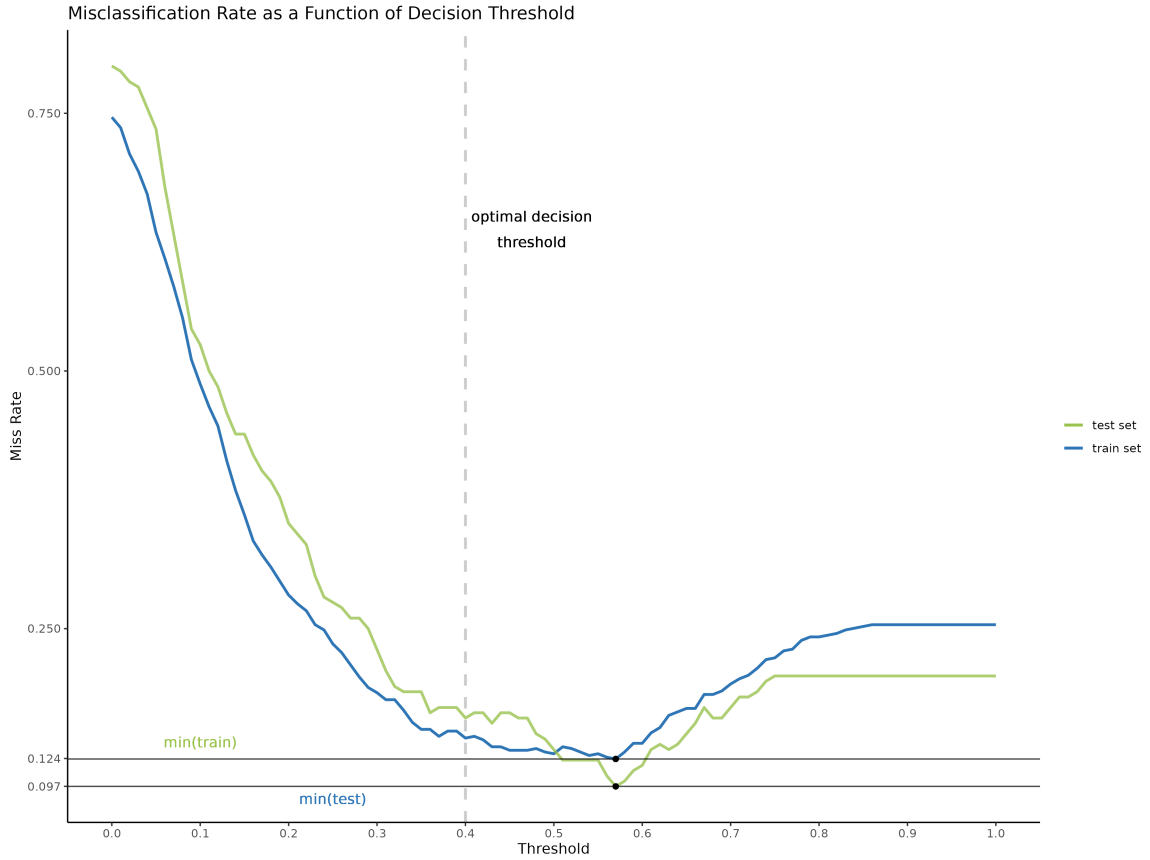


Table A.X: C2C Discretionary Arrest Outcomes, Ever Arrested Drug/Other

			ITT		TOT		
Outcome	N	Control Mean	Estimate	P-Value	CCM	Estimate	P-Value
Drug Arrests (High Discretion)							
6 Months	2074	0.008	0.013***	0.008	0.000	0.022***	0.007
				(0.005)			(0.008)
12 Months	2074	0.017	0.009	0.140	0.000	0.015	0.134
				(0.006)			(0.010)
24 Months	2074	0.031	0.005	0.500	0.011	0.009	0.493
				(0.008)			(0.012)
36 Months	2074	0.043	0.006	0.496	0.020	0.010	0.489
				(0.009)			(0.014)
Drug Arrests (Low Discretion)							
6 Months	2074	0.004	0.000	0.851	0.004	-0.001	0.848
				(0.003)			(0.004)
12 Months	2074	0.011	-0.007*	0.067	0.016	-0.011*	0.064
				(0.004)			(0.006)
24 Months	2074	0.017	-0.006	0.267	0.018	-0.009	0.260
				(0.005)			(0.008)
36 Months	2074	0.023	-0.009	0.115	0.029	-0.015	0.110
				(0.006)			(0.009)
Other Arrests (High Discretion)							
6 Months	2074	0.034	-0.009	0.197	0.030	-0.015	0.190
				(0.007)			(0.011)
12 Months	2074	0.059	-0.007	0.451	0.042	-0.011	0.444
				(0.009)			(0.015)
24 Months	2074	0.097	-0.008	0.467	0.069	-0.013	0.459
				(0.011)			(0.018)
36 Months	2074	0.142	-0.022*	0.089	0.123	-0.035*	0.085
				(0.013)			(0.020)
Other Arrests (Low Discretion)							
6 Months	2074	0.062	-0.006	0.559	0.048	-0.009	0.553
				(0.010)			(0.015)
12 Months	2074	0.104	-0.023**	0.042	0.100	-0.038**	0.040
				(0.012)			(0.018)
24 Months	2074	0.155	-0.020	0.151	0.139	-0.032	0.145
				(0.014)			(0.022)
36 Months	2074	0.184	-0.023	0.117	0.168	-0.037	0.112
				(0.015)			(0.023)

ITT/TOT estimates were calculated using randomization strata, cohort level fixed effects, and robust standard errors. TOT estimates were computed using a 2SLS regression, where the participation rate was instrumented by the treatment random assignment. CCM is the control complier mean–those who would have taken up treatment had they been offered it. It is calculated by taking the mean of the outcome for those that comply with the treatment minus the TOT. We include demographic covariates (age/race/gender dummies), school grade, prior enrollment in free/reduced lunch benefits, an ever NOT enrolled in school at baseline indicator, and prior arrest records (including prior discretionary/non-discretionary arrest indicators).

Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

A.1.6 More details on CPD Stops

A.1.6.1 Matching to CPD Stop Data

To match our study participants to Chicago Police Department (CPD) stops data, we followed a similar but more nuanced protocol outlined in the Matching to CPD arrest data section. Additional criteria was required to determine a good match to CPD for many cases, as a large proportion of traffic and pre-2016 stops had incomplete date of birth fields (some traffic stops only include the birth-year of the person being stopped, and many pre-2016 have no information on birth date). Because there are no unique identifiers in the stops data that link one person to multiple stops (such as IR number in the CPD arrest data), we created a pseudo-identifier that combined an individual's first name, last name, and birth date to match to study participants. Similar to the arrest matching protocol, if name and DOB or name and address are an exact match, we kept the match (these account for nearly 70% of our matches). For cases in the stops data with complete DOB fields that did not match perfectly to a C2C participant, we used the same probabilistic matching criteria outlined in the arrest matching section, while also considering the potential match's race and gender to reduce the number of manual reviews the team would consider. These probabilistic matches account for approximately 16% of the C2C matches to the stops data.

For cases which a stop record only had birth-year, we required a higher standard to keep a potential match. Given the lack of available information to match on, we only considered potential matches that had the same birth-year. Matches with a high name-similarity score, a high address-similarly score, matching gender, and which did not have a common name in the CPD data are kept. Other matches with a high name-similarity score were manually reviewed. Matches with birth-year only account for approximately 12% of our matches. For cases with completely missing DOB information, we had to rely on the age field recorded in the stops data and compare it to the implied age of a given study participant at the date of the stop. Because the age field in the stops data might be a guess by the CPD officer, we considered all matches with a recorded age being within one year of a given study participant's real age on the date of the stop. We only keep matches with a missing DOB field if the name matches perfectly, the address matches perfectly with a high name-similarity score, or a very high name-similarly score with a name that is unique to the CPD data (1 or 0 unique IR numbers in the arrest data). These cases make up only 2% of our matches.

A.1.6.2 What Constitutes a Stop

CPD officially defines an investigative stop as non-voluntary contact. Specifically, guidance for officers states, that "a voluntary contact is a consensual encounter between an officer and

a person during which the person must feel free to leave the officer's presence. An officer may approach any person at any time for any reason on any basis. However, absent reasonable suspicion or probable cause, that person must be free to walk away at any time. An officer's ability to articulate that no factors existed that would make a reasonable person perceive they were not free to leave is important. The following are some factors the court may consider to determine whether or not a consensual encounter has elevated to an Investigatory Stop or an arrest: Threatening presence of several officers, Display of a weapon by an officer; Use of language or tone of voice indicating that compliance with the officer's request might be compelled; Officer blocks a person's path; or Choice to end the encounter is not available to the person." Officers are required to complete ISR reports, which are quite lengthy and require detailed characteristics about the person and the reason or factors that led to the stop intended to demonstrate reasonable, articulable suspicion. They also detail whether or not a protective pat down was conducted, the reasons for the pat down, and if and why a subsequent search was then done. A narrative field also details the factors and situation of the stop. Our stops data includes every stop (traffic or street) that was formally recorded by CPD, even if that stop resulted in an arrest. In cases where an arrest came from a stop, these incidents are recorded in both the stops and the arrest data.

For traffic stops, police officers in Chicago are required to fill out traffic stop data sheet that specifies the driver's name, race, the reason for the stop, whether or not a search was conducted, whether or not contraband was found during the search, and the action resulted.[20a]

[20a]Since 2004, given concern around the racial disparities in stops, the Illinois Traffic and Pedestrian Stop Statistical Study Act became law and required all Illinois law enforcement to document the race of the driver and the reason for the stop, and report traffic stops to the Illinois Department of Transportation. Illinois, and subsequently Chicago, has some of the most rigorous data collection requirements around traffic stops in the nation.

Table A.XI: C2C Stop Outcomes, Ever Stopped and Number of Stops

		ITT		TOT		
Outcome	Control Mean	Estimate	P-Value	CCM	Estimate	P-Value
Ever Stopped						
6 months	0.218	0.009	0.593	0.158	0.014	0.587
		(0.016)			(0.026)	
12 months	0.307	0.027	0.127	0.236	0.044	0.121
		(0.018)			(0.029)	
24 months	0.446	0.036	0.060*	0.391	0.059	0.056*
		(0.019)			(0.031)	
36 months	0.546	0.010	0.607	0.512	0.016	0.601
		(0.020)			(0.031)	
Ever Stopped (Street)						
6 months	0.208	-0.003	0.869	0.163	-0.004	0.867
		(0.016)			(0.025)	
12 months	0.287	0.024	0.161	0.219	0.039	0.154
		(0.017)			(0.028)	
24 months	0.398	0.021	0.272	0.350	0.033	0.264
		(0.019)			(0.030)	
36 months	0.471	0.008	0.683	0.433	0.013	0.678
		(0.019)			(0.030)	
Ever Stopped (Traffic)						
6 months	0.027	0.008	0.304	0.010	0.013	0.297
		(0.008)			(0.012)	
12 months	0.052	0.003	0.788	0.036	0.004	0.784
		(0.009)			(0.015)	
24 months	0.124	0.014	0.339	0.092	0.022	0.331
		(0.014)			(0.023)	
36 months	0.212	-0.008	0.662	0.197	-0.012	0.656
		(0.017)			(0.027)	
Number of Stops						
6 months	0.437	-0.020	0.634	0.315	-0.032	0.628
		(0.041)			(0.066)	
12 months	0.850	-0.025	0.732	0.579	-0.040	0.728
		(0.072)			(0.115)	
24 months	1.696	-0.017	0.894	1.333	-0.028	0.892
		(0.129)			(0.206)	
36 months	2.642	-0.062	0.741	2.209	-0.101	0.737
		(0.188)			(0.300)	
Number of Street Stops						
6 months	0.393	-0.019	0.616	0.280	-0.031	0.610
		(0.038)			(0.060)	
12 months	0.753	-0.023	0.719	0.511	-0.038	0.714
		(0.064)			(0.103)	
24 months	1.363	-0.011	0.917	1.021	-0.018	0.916
		(0.106)			(0.170)	
36 months	1.945	-0.069	0.624	1.530	-0.112	0.618
		(0.141)			(0.225)	
Number of Traffic Stops						
6 months	0.044	-0.001	0.963	0.035	-0.001	0.963
		(0.014)			(0.022)	
12 months	0.097	-0.002	0.952	0.069	-0.003	0.951
		(0.026)			(0.041)	
24 months	0.333	-0.006	0.914	0.312	-0.010	0.913
		(0.058)			(0.093)	
36 months	0.697	0.007	0.943	0.679	0.012	0.942
		(0.100)			(0.160)	

Notes: CM is the control mean. Intent-to-treat (ITT) and Treatment on the treated (TOT) estimates were calculated using randomization block fixed effects and robust standard errors. CCM is the control complier mean. We include the following baseline characteristics: demographic covariates (age/race/gender dummies), school grade at randomization indicators, prior enrollment in free/reduced lunch benefits, prior gap in enrollment indicator, prior arrest records by type (numbers and indicators), indicator for any prior victimization, and number of prior victimizations. Standard errors are shown in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A.XII: Heterogeneous ITT Effects by Subgroup, Main Arrest Outcomes (24 months)

	Any Arrest	Number of Arrests	Any Violent Arrest	Number of Violent Arrests
Subgroup = Any baseline arrest				
Treat	-0.068***	-0.157***	-0.026*	-0.032
	(0.019)	(0.051)	(0.014)	(0.020)
Any bl arrest x Treat	0.010	0.286*	-0.031	-0.026
	(0.038)	(0.170)	(0.034)	(0.058)
CM (Any bl arrest = YES)	0.625	1.773	0.312	0.438
CM (Any bl arrest = NO)	0.190	0.370	0.082	0.108
Subgroup = Any violent baseline arrest				
Treat	-0.068***	-0.165***	-0.026*	-0.037*
	(0.019)	(0.055)	(0.014)	(0.020)
Any bl violent arrest x Treat	0.019	0.506**	-0.048	-0.017
	(0.044)	(0.225)	(0.043)	(0.081)
CM (Any bl violent arrest = YES)	0.671	1.957	0.346	0.498
CM (Any bl violent arrest = NO)	0.250	0.554	0.111	0.147
Subgroup = Enrollment gap at baseline				
Treat	-0.071***	-0.054	-0.040***	-0.040*
	(0.019)	(0.067)	(0.015)	(0.024)
Enrollment gap at bl x Treat	0.040	-0.004	0.016	-0.003
	(0.043)	(0.189)	(0.039)	(0.066)
CM (Enrollment gap at bl = YES)	0.461	1.432	0.218	0.325
CM (Enrollment gap at bl = NO)	0.316	0.729	0.151	0.201

Standard errors are robust; *** p<0.01, ** p<0.05, * p<0.1

Notes: The number of participants that had an arrest at baseline is N = 728 (35.1% of the sample). The number of participants that had an arrest associated to a violent charge at baseline is N = 458 (22.1% of the sample). The number of participants that had a school enrollment gap at baseline is N = 412 (19.9% of the sample).

Table A.XIII: Heterogeneous ITT Effects by Subgroup, Discretionary Arrests Outcomes (24 months)

	Any High Discretion Arrest	Number of High Discretion Arrests	Any Low Discretion Arrest	Number of Low Discretion Arrests
Subgroup = Any baseline arrest				
Treat	-0.015	-0.028	-0.068***	-0.128***
	(0.010)	(0.017)	(0.018)	(0.043)
Any bl arrest x Treat	0.008	0.090	0.040	0.177
	(0.032)	(0.072)	(0.039)	(0.137)
CM (Any bl arrest = YES)	0.296	0.488	0.545	1.285
CM (Any bl arrest = NO)	0.043	0.059	0.174	0.311
Subgroup = Any violent baseline arrest				
Treat	-0.027**	-0.048**	-0.062***	-0.122***
	(0.011)	(0.023)	(0.018)	(0.047)
Any bl violent arrest x Treat	0.067	0.234***	0.032	0.251
	(0.041)	(0.090)	(0.046)	(0.184)
CM (Any bl violent arrest = YES)	0.329	0.511	0.597	1.446
CM (Any bl violent arrest = NO)	0.076	0.125	0.221	0.429
Subgroup = Enrollment gap at baseline				
Treat	-0.013	0.002	-0.063***	-0.064
	(0.013)	(0.024)	(0.019)	(0.056)
Enrollment gap at bl x Treat	0.001	0.008	0.040	-0.014
	(0.038)	(0.096)	(0.044)	(0.148)
CM (Enrollment gap at bl = YES)	0.243	0.447	0.388	0.985
CM (Enrollment gap at bl = NO)	0.105	0.153	0.286	0.576

Standard errors are robust; *** p<0.01, ** p<0.05, * p<0.1

Notes: The number of participants that had an arrest at baseline is N = 728 (35.1% of the sample). The number of participants that had an arrest associated to a violent charge at baseline is N = 458 (22.1% of the sample). The number of participants that had a school enrollment gap at baseline is N = 412 (19.9% of the sample).

A.1.7 Incarceration

If youth are accumulating violent arrests as well as other significant charges, they may be more likely to be incarcerated (and for long periods of times) and therefore unable to be stopped or arrested by police officers. Our findings may be difficult to interpret if there are substantially more control youth incapacitated compared to treatment youth (and therefore

more present in the neighborhood, etc). First, we want to highlight that incarceration rates among youth have fallen substantially in the last 5-10 years in Chicago and Illinois more broadly due to concerted policy efforts. For example, in July 2019 there were 264 youth detained in the Illinois Department of Juvenile Justice (IDJJ) centers across the state. [21a] In Chicago, the Juvenile Temporary Detention Center (JTDC) of Cook County on average houses roughly 100 youth at any given point in time, with stays averaging around a week or two.

Despite the general downward trends, we also attempt to confirm incarceration rates among our study youth. Although we don't have juvenile incarceration data available, we can use CPS data as a proxy for incarceration. CPS records leave reasons for youth, including the reason of being legally committed to a correctional institution. CPS also has indicators for whether or not youth are attending the alternative schools available to detained students.[22a] We use the combination of these indicators, as well as publicly available adult incarceration data from the Illinois Department of Corrections (IDOC) as a combined measure of being incarcerated post randomization. In Table A.III we present the RCT effects looking at incarceration as an outcome up to 18 months post randomization. We find that the base rates are small (about 18 days incarcerated in 18 months), and there is no differential impact between treatment and control youth in almost any outcome period. This suggests that incapacitation effects due to incarceration will not change the interpretation of our findings. We also combine this incarceration measure with transfer out of Chicago data to create an indicator for any form of data censoring, and generally find no differential data censoring in our study sample (see Table A.III). As the youth age, we are more likely to observe C2C study youth in adult incarceration data over time. A complement paper will present these longer term incarceration results.

[21a] `https://www2.illinois.gov/idjj/Pages/Data-and-Reports.aspx`

[22a] There are two CPS schools that serve students that are detained. Nancy B Jefferson Alternative High School serves youth in the JTDC. While York Alternative High School works with students 17 and older who are detained in the Cook County Jail.