UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating the Short-Term Impact of Gang Prevention Services on a Sample of Los Angeles Youth using Local Randomization Regression Discontinuity Design

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Georgina Garcia-Obledo

2020

ABSTRACT OF THE THESIS

Evaluating the Short-Term Impact of Gang Prevention Services on a Sample of Los Angeles Youth using Local Randomization Regression Discontinuity Design

by

Georgina Garcia-Obledo
Master of Science in Statistics
University of California, Los Angeles, 2020
Professor Erin K. Hartman, Chair

Since 2009, the City of Los Angeles Mayor's Office of Gang Reduction and Youth Development (GRYD) has provided gang prevention services to at-risk youth ages 10 to 15 throughout the city. Youth are assigned to secondary prevention (full program) or primary prevention (reduced program) based on a score computed from the YSET, an attitudinal and behavioral assessment developed by GRYD. The same assessment is repeated periodically to track client progress. In this paper, we employ a local randomization regression discontinuity design (RDD) to assess the causal intention-to-treat (ITT) effect on retest YSET responses for a matched sample of barely eligible and barely ineligible individuals. We fail to find evidence of an ITT effect on most outcomes tested. Our analysis relies on strong assumptions and is limited by the available data. We believe additional data are essential for better evaluation of the impact of this extensive city-wide program.

The thesis of Georgina Garcia-Obledo is approved.

Paul Jeffrey Brantingham

Frederic R. Paik Schoenberg

Erin K. Hartman, Committee Chair

University of California, Los Angeles

2020

To my parents, Blas and Araceli, and my sister, Paola.

To everyone dedicated to youth development and violence prevention.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This research was funded by the City of Los Angeles contract number C-132202 ("GRYD Research and Evaluation"). Permission to use these data was provided by the City of Los Angeles Mayor's Office of Gang Reduction and Youth Development (GRYD). Any opinions, findings, conclusions or recommendations expressed in this study, however, are those of the author(s) and do not necessarily reflect the views of the GRYD Office. The YSET was created by the City of Los Angeles Mayor's Office of Gang Reduction and Youth Development and is the copyright of the City of Los Angeles.

I am greatly indebted to my awesome thesis committee. Professor Brantingham, thank you for the opportunity to partake in this research and for meeting with me multiple times a month to answer my questions and keep me on track. Professor Hartman, thank you for all of the practical tips, words of encouragement, and thorough answers to my many questions. Professor Schoenberg, thank you for the quick responses and helpful suggestions. I would also like to acknowledge Professor Hazlett's valuable guidance on causal inference and study design.

I would like to thank the many people who helped me get to graduate school and have supported me over the past two years. First and foremost, thank you to my wonderful parents whose unconditional love strengthens me every day. To my amazing sister, thank you for the constant encouragement from day one and the vital virtual handholding while I was writing my thesis. To my professors at Occidental College, thank you for inspiring me to continue my education. To the statistics department, thank you for all the patience and support. In particular, I would like to thank Professors Christou, Fletcher, Handcock, and Hazlett for their enthusiasm, generosity, and understanding. I would also like to thank my TA Mojtaba Sahraee-Ardakan for helping me improve and build confidence in my coding. Glenda Jones, thank you for making me feel supported and cared for from the day I first set foot on the eighth floor of the Math Sciences Building. Chie Ryu, Laurie Leyden, Jordan Price, and Marissa Martinez, thank you for the administrative support. The Graduate Writing Center's workshops, one-on-one appointments with Jesslyn Whittell, and weekly writing hours have

made the writing process more manageable. Thank you to the Gates Millennium Scholarship Program for providing financial support, learning opportunities, and an amazing community for the past nine years.

Finally, I would like to extend my gratitude to my friends, old and new, whose companionship and assistance has been essential. Audrey Larkin, thank you for being my UCLA tour guide, helping me when I was injured, and providing feedback on my thesis. Mikhail (Misha) Vysotskiy, thank you for finding a funny typo in a previous version of this paper. Vicky Kelley, thank you for the innumerable FaceTime and Zoom study sessions. Alicia, Annie, Levina, Ricardo, Sarah, and Victoria, thank you for your encouragement. To my kind classmates, especially Alex, Conor, Helen, Pablo, Pedro, Onyambu, Shuchi, and Thomas, thank you for helping me with schoolwork and graduate school life in general. To everyone else who has provided academic, physical, and moral support throughout this process, thank you.

CHAPTER 1

Introduction

In 2009, the City of Los Angeles Mayor's Office of Gang Reduction and Youth Development (GRYD) contractors began providing gang prevention services in various Los Angeles neighborhoods. Two years later, these services were standardized in accordance with the GRYD Prevention Services model. Thousands of Los Angeles youth have participated in the GRYD prevention programs offered in over twenty designed geographic areas throughout the city. There are two levels of programming offered. The full version, called "secondary prevention" or "model services," consists of individual and family sessions and group activities. One program cycle is designed to last approximately six months, and clients may continue for additional cycles as needed. The reduced version, called "primary prevention," is available for individuals who do not qualify for the full program (Kraus et al., 2017).

Eligibility for secondary prevention is determined by three factors. First, clients must be 10 to 15 years old. Second, clients must "have a significant presence in a GRYD [geographic] Zone" (p. 4). Finally, youth must have a risk score of 4 or higher on the Youth Services Eligibility Tool (YSET) assessment that is administered upon referral to GRYD Prevention Services.[1] If the risk score is 3 or lower, the individual is eligible for primary prevention.[2] Six months after enrollment or upon the completion of a program cycle (whichever comes first), the YSET is administered again. At this point, individuals may enroll for another cycle, depending on their progress (Kraus et al., 2017).

Despite the large amount of data that have been collected over the years, causal inference on the effect of GRYD Prevention Services is limited by the fact that treatment assignment

[1]The YSET is the copyright of the City of Los Angeles (2012). All rights reserved.

[2]These treatment group assignments are not final. See Chapter 2 for more details.

is not random. In this paper, we use a local randomization regression discontinuity design (RDD) for causal inference. We make use of the known treatment assignment rule to evaluate the intention-to-treat (ITT) effect of secondary prevention (as opposed to primary prevention) on retest YSET responses. After cleaning the data, we use matching to identify a sample of barely eligible and barely ineligible individuals that are balanced on age, intake year, race/ethnicity, and gender. We then use standard randomization inference tools to obtain p-values and confidence intervals for the difference-in-means of the overall risk score and component scale scores at retest for this sample. We do not find evidence of a significant difference between barely eligible and barely ineligible youth in overall risk score or most component scales at retest. However, these results depend on strong assumptions and on finite-sample randomization inference and so may not generalize broadly.

In Chapter 2, we describe the YSET in more detail and introduce the notation and potential outcomes framework we use throughout the paper. We also specify the subset of data used for analysis. Next, in Chapter 4, we formally introduce the fundamental assumptions of local randomization regression discontinuity design and the estimation and inference methods we use in our analysis. Chapter 5 describes falsification tests and presents covariate balance test results within our chosen local randomization window to demonstrate why we pre-process our data by matching. In Chapter 6, we specify the matching process used to isolate a sample for which the local randomization RDD identifying assumptions are plausible. We present the results of our analysis in Chapter 7. Chapter 8 concludes the paper with a discussion about the takeaways and limitations of our analysis. We also describe further work that can be done with the existing data and with additional data to test for sensitivity and robustness and to better evaluate GRYD Prevention Services. Most figures are in Appendix A.

CHAPTER 2

Data, Notation, and Potential Outcomes Framework

GRYD maintains a database of all Youth Service Eligibility Tool (YSET) interviews conducted since the program inception, regardless of whether individuals enrolled in Prevention Services or not. The YSET, developed by GRYD, is used to determine eligibility for secondary prevention. Since our analysis is based on the YSET, we will now describe this assessment in further detail. We will also introduce the notation we used throughout the paper and our data cleaning process.

2.1 YSET Scoring

The YSET can be divided into two parts. The first part consists of demographic and administrative questions. The second part contains attitudinal, behavioral, and family questions, divided into several themed sets, called "scales." Most questions are multiple choice or yes/no questions. Each scale has an integer score Y_{iSt}, where $i \in \{1, 2, ..., N\}$ is the case index, S is the scale label, and $t \in \{0, 1, ...\}$ is the YSET number ($t = 0$ = intake YSET, $t = 1$ = first retest YSET, and so on). Nine of these scales (A, B, C, DE, F, G, H, IJ, and T) are used to compute the risk score that determines eligibility for secondary prevention. For each of these nine scales, there is a pre-determined value r_S at or above which scores are considered "concerning" or indicative of risk. Table 2.1 lists the scales used to compute the risk score, along with their descriptions, raw score ranges,[1] and thresholds (GRYD, 2013).

The risk score $X_{it} \in \{0, 1, ..., 9\}$ is the number of scale scores that are concerning.

[1]These are the ranges if all questions are answered. We kept observations with scale raw scores below the minimum score even though this implies that some questions were not answered.

Scale	Description	Raw Score Y_{iSt}	Risk Factor Threshold (Concerning If $Y_{iSt} \geq r_S$)
A	Antisocial Tendencies	$Y_{iAt} \in \{5, 6, ..., 30\}$	$R_{iAt} = \mathbb{1}(Y_{iAt} \geq 16)$
B	Weak Parental Supervision	$Y_{iBt} \in \{3, 4, ..., 15\}$	$R_{iBt} = \mathbb{1}(Y_{iBt} \geq 7)$
C	Critical Life Events	$Y_{iCt} \in \{0, 1, ..., 7\}$	$R_{iCt} = \mathbb{1}(Y_{iCt} \geq 4)$
DE	Impulsive Risk Taking	$Y_{iDEt} \in \{4, 5, ..., 20\}$	$R_{iDEt} = \mathbb{1}(Y_{iDEt} \geq 14)$
F	Guilt Neutralization	$Y_{iFt} \in \{5, 6, ..., 30\}$	$R_{iFt} = \mathbb{1}(Y_{iFt} \geq 19)$
G	Negative Peer Influence	$Y_{iGt} \in \{5, 6, ..., 25\}$	$R_{iGt} = \mathbb{1}(Y_{iGt} \geq 13)$
H	Peer Delinquency	$Y_{iHt} \in \{5, 6, ..., 30\}$	$R_{iHt} = \mathbb{1}(Y_{iHt} \geq 12)$ $_{\text{If age}_{it}<13}$ $R_{iHt} = \mathbb{1}(Y_{iHt} \geq 14)$ $_{\text{If age}_{it}\geq 13}$
IJ	Self-Reported Delinquency and Substance Use	$Y_{iIJt} \in \{0, 1, ..., 17\}$	$R_{iIJt} = \mathbb{1}(Y_{iIJt} \geq 4)$ $_{\text{If age}_{it}<13}$ $R_{iIJt} = \mathbb{1}(Y_{iIJt} \geq 6)$ $_{\text{If age}_{it}\geq 13}$
T	Family Gang Influence	$Y_{iTt} \in \{0, 1, 2\}$	$R_{iTt} = \mathbb{1}(Y_{iTt} \geq 1)$

Table 2.1: Nine Scales Used to Compute Risk Score X_{it} (January 2013 Instructions)

Mathematically, it is the sum of nine indicator functions,

$$X_{it} = R_{iAt} + R_{iBt} + R_{iCt} + R_{iDEt} + R_{iFt} + R_{iGt} + R_{iHt} + R_{iIJt} + R_{iTt},$$

where

$$R_{iSt} = \mathbb{1}(Y_{iSt} \geq r_S) = \begin{cases} 1 & \text{if } Y_{iSt} \geq r_S \\ 0 & \text{if } Y_{iSt} < r_S. \end{cases}$$

If $X_{it} \geq 4$, then individual i is eligible for secondary prevention based on YSET t, given that i is 10 to 15 years old and "[has] a significant presence in a GRYD [geographic] Zone" (Kraus et al., 2017, p. 4). Otherwise, individual i is not eligible for secondary prevention but is eligible for primary prevention, assuming she meets the two other criteria. Initial treatment assignment based on the YSET risk score (i.e. eligibility) does not necessarily equal the "final" treatment assignment or even the actual treatment received. If an individual's intake risk score is below 4, but the interviewer believes secondary prevention is more suitable based on the in-person assessment, the site can file a formal assignment challenge with GRYD. The challenge can result in the individual being assigned to secondary prevention even though her risk score is below 4. If the interviewer suspects gang involvement, the site formally consults with GRYD. The consultation determines if Prevention Services or Intervention Services (or

a combination of both) are better suited for the individual. In addition, there are several cases in the YSET database where an individual is enrolled in secondary prevention despite an intake risk score below 4 and no documented assignment challenge. There are also a handful of cases where the individual enrolls in primary prevention despite being eligible for secondary prevention. That is, there is two-sided noncompliance with initial treatment assignment, and we do not always know why some clients' actual enrollment differs from their initial treatment assignment. However, in the vast majority of cases, clients enroll in the treatment level to which they were initially assigned.

2.2 Potential Outcomes Framework

In this paper, we use values from the retest YSET as our outcomes. Of course, since the ultimate goal of Prevention Services is to keep youth from joining gangs, it would be ideal (from a program evaluation perspective) to know whether or not individuals enrolled in the programs eventually join gangs. We do not have these data, and collecting such data presents serious challenges and concerns. However, the YSET was carefully designed to identify risk factors for joining a gang, and its predictive value was validated by a 2015 study (Hennigan, Kolnick, Vindel, and Maxson). One of the goals of GYRD Prevention Services is to decrease scale scores and the overall risk score (Kraus et al., 2017). Therefore, inference on retest YSET outcomes does provide information on the short-term effects of GRYD that may translate to long-term behavioral effects. The credibility of the point estimates and p-values presented here depends on the extent to which we believe that the identifying assumptions described in Chapters 4 and 6 are realistic. The generalizability to individuals far from the eligibility cutoff also depends on the extent to which we believe the study sample is representative of the population of youth enrolled in GRYD Prevention Services.

We utilize the Neyman-Rubin potential outcomes framework for causal inference (Rubin, 1974, 1986; Imbens & Rubin, 2015; Sekhon, 2009). We will be using notation similar to that in Imbens and Rubin (2015) and Cattaneo et al. (2019). This framework allows us to clearly specify the causal effect, at least theoretically. In order to specify potential outcomes

with this notation, we must first specify the notation for initial treatment assignment and actual treatment enrollment level indicators. We must specify both because there is two-sided non-compliance. We will use Z_{it} for the initial treatment assignment based on YSET t and D_{it} for the actual treatment enrollment level between YSET t and YSET $t+1$. More specifically,

$$
\begin{aligned}
Z_{it} &= \mathbb{1}(X_{it} \geq 4) \\
&= \begin{cases} 1 & \text{if } X_{it} \in \{4,5,6,7,8,9\} \\ 0 & \text{if } X_{it} \in \{0,1,2,3\} \end{cases} \\
&= \begin{cases} 1 & \text{if client } i \text{ is assigned to secondary prevention/} \\ & \text{is eligible for secondary prevention} \\ & \text{based on YSET } t \\ \\ 0 & \text{if client } i \text{ is assigned to primary prevention/} \\ & \text{is } \underline{\text{in}}\text{eligible for secondary prevention} \\ & \text{based on YSET } t \end{cases}
\end{aligned}
$$

and,

$$
D_{it} = \begin{cases} 1 & \text{if client } i \text{ is enrolled in secondary prevention} \\ & \text{between YSET } t \text{ and YSET } t+1 \\ \\ 0 & \text{if client } i \text{ is enrolled in primary prevention} \\ & \text{between YSET } t \text{ and YSET } t+1. \end{cases}
$$

We make the standard Stable Unit Treatment Value Assumption (SUTVA), which translates to the assumption that there is no interference between units and that each client only receives one version of treatment at each level (Rubin, 1980; Imbens & Rubin, 2015). Note that this does not mean that secondary prevention and primary prevention are exactly the same for all individuals. Rather, this means that if an individual is assigned to secondary prevention, there is one "fixed" customized version of secondary prevention for that client and that if the individual is assigned to primary prevention, there is one "fixed" customized version of primary prevention for her. The fact that Prevention Services are customized

(within bounds) to each client does not invalidate the SUTVA assumption. Furthermore, this assumption means that client i's eligibility and treatment enrollment level are independent of client j's eligibility and treatment enrollment level, $i \neq j$. That is Z_{it}, $D_{it} \perp Z_{jt}$, D_{jt}, $i \neq j$. This assumption simplifies estimation. We believe this assumption holds in general, although there are cases where two or more siblings enroll in the program, and this might mean that they all receive the same level of treatment even if it does not correspond to all of their individual intake risk scores.

For our primary analysis, we will focus on the "intention-to-treat" (ITT) effect, i.e. the effect of the initial treatment assignment on retest responses, which we can analyze with a "sharp" local randomization regression discontinuity design (RDD). In order to account for non-compliance, we could use a "fuzzy" local randomization RDD, but since compliance rates are high, we do not anticipate that the corresponding point estimates and inference would be significantly different. The intention-to-treat (ITT) causal effect of interest is defined as $\tau_i^{ITT} = Y_{i1}(Z_{i0} = 1) - Y_{i1}(Z_{i0} = 0)$. In other words, τ_i^{ITT} is the causal effect of eligibility as opposed to ineligibility determined by i's intake YSET on client i's first retest YSET. Unfortunately, we cannot actually identify the causal effect on an individual basis. This is because we only observe one of $Y_{i1}(Z_{i0} = 1)$ and $Y_{i1}(Z_{i0} = 0)$. However, we can estimate causal effects by averaging over various treatment and control observations (Imbens & Rubin, 2015).

2.3 Data Cleaning

The YSET database contains anonymized YSETs for over 22,000 unique cases from 2009 to 2020.[2] However, only a fraction have at least two YSETs with valid scores for all nine

[2]Each case is assigned an ID number. The same person may have various ID numbers, as they are assigned at each intake. For instance, if an individual completes a cycle of primary prevention and then switches to secondary prevention based on her retest, she will be assigned a new ID number. If an individual leaves the program and later returns, she will also be assigned a different ID number. The intake number is documented, so we are able to identify individuals without previous direct exposure to the program. Unfortunately, there are some cases where one ID number appears to correspond to more than one individual. We drop cases where there are two or more demographic inconsistencies between intake and retest YSETs since that suggests that the ID corresponds to multiple individuals.

scales used for scoring.[3] We limit our analysis to cases where the earliest YSET is listed as a first intake and the following YSET is listed as a first retest.[4] Following Kraus et al. (2017), we only consider cases where the retest YSET is administered four to eight months after the intake interview. The retest is supposed to happen after the completion of a GRYD cycle or six months after enrollment, whichever comes first (Kraus et al., 2017). In deciding which cases to keep in our dataset, we use the intake interview date rather than the enrollment date as the start date because we believe this makes potential outcomes more comparable, given the sometimes rapid nature of adolescent attitudinal and behavioral development. Even though this means dropping many observations, we believe that doing so will reduce bias; considering cases with time differences between intake and retest outside of the four to eight month range could be a source of additional confounding because there may have been more issues with attendance, etc. in the cases with more time between tests. Our final dataset contains 5,298 unique IDs. From now on, we will assume that each ID corresponds to a distinct individual, so we will be using the terms ID, case, individual, and client interchangeably.

[3]Table 2.1 lists the scale score ranges if all questions are answered. Scores below these ranges imply that at least one question was not answered, but we keep these observations anyway.

[4]There are no individuals in our final dataset that were dismissed for failing to complete the program and/or long-term non-attendance, according to an additional administrative file.

CHAPTER 3

Previous Evaluations

With a program as extensive as GRYD Prevention Services, evaluation is extremely important. Since GRYD began providing prevention services in 2009, there have been various evaluations of the program. In 2013, the GRYD annual evaluation report used a continuity-based fuzzy regression discontinuity design approach. The report used changes in raw scores as outcomes and found evidence supporting the benefits of GRYD Prevention Services (Dunworth et al., 2013). Since then, local randomization regression discontinuity designs using finite-sample valid randomization inference have been further developed in the literature. Due to the coarse and discrete nature of the running variable, X_{i0}, we believe a local randomization RDD approach is more appropriate for this dataset.

The 2017 Prevention Services evaluation report used a difference-in-differences model to compare the outcomes of individuals in secondary prevention with those of individuals in the greater Los Angeles area with similar intake YSET scores that were not referred to GRYD Prevention Services. The report found that the individuals in secondary prevention showed significantly more improvement on YSET scales than the control group (Kraus et al., 2017). We hope that the present study serves as a complement to these previous evaluations.

CHAPTER 4

Local Randomization Regression Discontinuity Design Assumptions and Estimation

4.1 RDD Theory and Assumptions

Experiments are the ideal approach for studying causality, but they are often infeasible and/or unethical to conduct, particularly in the social sciences. Various "quasi-experimental" strategies have been developed for "extracting" experiments from observational data. Regression discontinuity design (RDD) is one such strategy. The basic idea underlying RDD is that when treatment is deterministically assigned based on some numerical score with a cutoff, the individuals close to the cutoff on either side may not be terribly different. That is, near the cutoff, treatment assignment is somewhat arbitrary, as it would be in an experimental setting. In 1960, Thistlethwaite and Campbell realized that the somewhat arbitrary nature of a treatment cutoff could be used to analyze the causal effect of a program or policy. Since then, RDD has been developed in various fields (Cook, 2008).

Regression discontinuity designs can be divided into two categories: continuity-based RDDs and local randomization RDDs. Continuity-based RDDs are more prevalent, but they are not suited to our data because our running variable, the intake risk score X_{i0}, is discrete and only takes on 10 possible values. On the other hand, local randomization RDDs can accommodate cases like ours, provided that certain assumptions are made. The key assumption for local randomization RDD is that treatment assignment is "as-if random" in a thin window encompassing the cutoff (Cattaneo, Titiunik, & Vazquez-Bare, 2017). Since the running variable is so coarse, we think that the only window in which this assumption would hold is the smallest possible window, consisting of those who are barely ineligible ($X_{i0} = 3$)

and those who are barely eligible ($X_{i0} = 4$). In other words, we assume that the potential outcomes at retest for those whose observed intake risk score was either 3 or 4 are independent of their eligibility based on the intake YSET. That is, $Y_{i1}(Z_{i0} = 1),\ Y_{i1}(Z_{i0} = 0) \perp X_{i0}$ for $X_{i0} \in \{3, 4\}$.

Implicit in the assumption that treatment assignment in the window $X_{i0} \in \{3, 4\}$ is "as-if random" is the assumption that there is no "*precise* manipulation" of the risk score in this window [emphasis added]. That is, there is a random component to X_{i0} (Lee, 2008). We believe this assumption is plausible because the complexity of the risk score calculation process would make it very difficult for potential clients to know exactly how to respond to guarantee their desired eligibility outcome. Clients may modulate their responses depending on whether or not they want to be eligible for secondary prevention, and interviewers may nudge or influence potential clients to answer questions a certain way, but such behavior does not constitute "precise manipulation" of the risk score.

Unfortunately, the local randomization assumption means that we cannot use local randomization RDD to analyze changes between intake and retest in the scale scores used to determine eligibility because these values necessarily depend on the values that determine the running variable, X_{i0}.[1] In order to analyze such changes, we could use a difference-in-differences approach. However, doing so would require the arguably stronger assumption of parallel trends and the use of data further from the cutoff. A difference-in-differences approach would imply assuming that data further from the cutoff on both sides are comparable, at least in terms of trends between intake and retest YSETs, which we believe is less plausible than the assumptions needed for local randomization RDD.

[1]Cattaneo et al. propose a way to use parametric polynomial transformations of the running variable as outcomes in a local randomization RDD (2017). However, the relationship between scale scores and the risk score is more complicated, so this approach does not apply.

4.2 Estimation and Inference

Under the assumption that treatment assignment is "as-if random" in the selected window around the cutoff, we can use traditional methods for analyzing experiments. We assume that potential outcomes are fixed and that the treatment assignment mechanism is known. This means that the only source of randomness is the treatment assignment. Under these assumptions, the observed difference-in-means is an unbiased point estimate of the average [intention-to-treat] treatment effect (ATE) (Imbens & Rubin, 2015).

We use randomization inference to calculate exact p-values testing the sharp null hypothesis of no treatment effect and to construct confidence intervals. The beauty of this framework is that we do not need to make any additional distributional assumptions about the outcomes or null distribution of our test statistic. In our analysis, we use the difference-in-means as our test statistic. The approaches described in this section are valid for inference about the clients whose data are used for computation, but the generalizability of the estimates and inference to cases outside of this sample is not guaranteed (Imbens & Rubin, 2015).

We assume that the number of units in the sample assigned to secondary prevention is fixed and that all possible permutations of the treatment assignment vector $\boldsymbol{Z_0}$ are equally likely. There are $\binom{N}{n_1}$ possible permutations of $\boldsymbol{Z_0}$, where $N = n_1 + n_0$ is the sample size, $n_1 = \sum_{i=1}^{N} Z_{i0}$ is the number of clients eligible for secondary prevention, and $n_0 = \sum_{i=1}^{N}(1 - Z_{i0})$ is the number of clients that are not. The sharp null hypothesis of no treatment effect is $H_0 : Y_{i1}(Z_{i0} = 1) - Y_{i1}(Z_{i0} = 0) = 0$ for $i = 1, 2, ..., N$, which means that the average treatment effect is also zero. The null distribution of the difference-in-means for this hypothesis can be constructed by computing the difference-in-means of each possible permutation of $\boldsymbol{Z_0}$. The Fisher exact p-value is the proportion of differences-in-means under the null hypothesis that are at least as extreme as the observed difference-in-means. Computing the difference-in-means for every permutation of $\boldsymbol{Z_0}$ is not practical for anything other than small values of N, since the number of permutations increases dramatically as N increases. However, the null distribution and thus the Fisher exact p-value can be estimated with randomly drawn

treatment assignment vectors. We will refer to these estimated Fisher exact p-values as "randomization p-values," following the example of Cattaneo, Titiunik, and Vazquez-Bare (2016).

The same procedure can be repeated for a range of sharp null hypotheses of the form $H_0 : Y_{i1}(Z_{i0} = 1) - Y_{i1}(Z_{i0} = 0) = \tau$ for $i = 1, 2, ..., N$ to compute a $1 - \alpha$ confidence interval: the confidence interval is bounded by the most extreme values of τ for which the null hypothesis is not rejected at a significance level of α (Imbens & Rubin, 2015). These hypothesis test inversion confidence intervals are sensitive to the values of τ tested. The function `rdrandinf` in the `rdlocrand R` package implements these randomization methods to estimate Fisher exact p-values and construct confidence intervals (Cattaneo et al., 2016).

CHAPTER 5

Falsification Tests

Before proceeding with the randomization inference, we evaluate the set of all clients with intake risk scores $X_{i0} \in \{3, 4\}$ in our cleaned dataset to determine whether the assumption that treatment assignment within the window is "as if random" is likely to hold. One way to test this assumption is to conduct balance tests on pre-treatment covariates (Lee, 2008; Cattaneo et al., 2017). If treatment assignment were truly random, we would expect pre-treatment covariates to be balanced between eligible and ineligible individuals. If the balance tests reveal significant imbalance, the credibility of the identifying assumption is brought into question.

We conduct balance tests using randomization inference on the difference-in-means of four pre-treatment covariates: age at intake, intake year, race/ethnicity, and gender. Bar charts showing the distributions of these four covariates are included in Appendix A. As shown in Table 5.1, there is significant imbalance in age at intake, intake year, and race/ethnicity. The average eligible client is older than the average ineligible client by a bit over two months; this difference has a randomization p-value of 0.012. That is, of the 10,000 difference-in-means calculated from randomly drawn, equally likely treatment assignment vectors, only 120 were as far from zero as the observed difference-in-means of 0.2 years. The average eligible client completed the intake YSET in 2015, while the average ineligible client completed the intake YSET in 2016; this difference has a randomization p-value of 0.000. The distribution of race/ethnicity is also significantly imbalanced. The proportion of eligible clients who are Black is significantly lower than the proportion of ineligible clients who are Black with a p-value of 0.001. The proportion of eligible clients who are Latinx is significantly higher than the proportion of ineligible clients who are Latinx with a p-value of 0.025. The proportion of

eligible clients who are multi-racial/ethnic is also significantly higher than that proportion of ineligible clients who are multi-racial/ethnic with a p-value of 0.094. There is not evidence of gender imbalance. However, the strong imbalance in the other covariates tested means that we cannot credibly use randomization inference on the entire set of barely eligible/ineligible clients in our cleaned dataset.

Another indication that the local randomization assumption may not hold is that there are many more barely eligible clients ($n_1 = 737$) than barely ineligible clients ($n_0 = 425$). Figure A.1 shows the distribution of all intake risk scores for the cleaned data. The jump in the distribution at the cutoff could be a sign of "sorting," i.e. manipulation of the risk score (Cattaneo et al., 2017). One way to address both concerns is to narrow down the sample using matching. Following the example of Keele, Titiunik, and Zubizarreta (2015), we will use matching to identify a sample for which the local randomization assumptions are more plausible.

Covariate	Barely Eligible ($n_1 = 737$)	Barely Ineligible ($n_0 = 425$)	Difference-in-Means	Randomization P-Value
Age at Intake	12.2	12.0	0.2	0.012**
Intake Year	2015.7	2016.7	-1.0	0.000***
Percent Asian	0.1	0.5	-0.4	0.560
Percent Black	17.1	25.6	-8.5	0.001***
Percent Latinx	77.7	71.8	5.9	0.025**
Percent Multi-Racial/Ethnic	3.4	1.6	1.8	0.094*
Percent Other Race/Ethnicity	0.8	0.2	0.6	0.270
Percent White	0.8	0.2	0.6	0.272
Percent Female	39.8	37.9	1.9	0.531
Note:				$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 5.1: Pre-Treatment Covariate Balance Checks Before Matching using Randomization Inference with 10,000 Permutations of the Eligibility Vector $\boldsymbol{Z_0} = (Z_{(1)0}, Z_{(2)0}, ..., Z_{(N)0})^T$

CHAPTER 6

Matching

Although we only have a few pre-treatment covariates to use in our balance tests, the strong imbalances indicate that the assumptions needed for local randomization are unlikely to hold for the set of all individuals with $X_{i0} \in \{3, 4\}$ in our cleaned dataset. However, if we can find a way to incorporate these covariates, we can more convincingly justify our RDD approach. Doing so requires the additional strong assumption of selection-on-observables, i.e. conditional ignorability, which we will make for the remainder of the analysis to hold (Keele et al., 2015).

Following Keele et al. (2015), we use matching to restrict our analysis to a balanced subset of observations. This decreases our sample size and further limits the generalizability of our results. However, we believe that it makes the local randomization assumptions more plausible. Matching is a way of pre-processing the data (Ho, Imai, King, & Stuart, 2007). Local randomization inference on the matched dataset is justified by the assumption that the potential outcomes under the two different initial treatment assignments of individuals with intake risk scores $X_{i0} \in \{3, 4\}$ are independent of X_{i0} conditional on age, intake year, race/ethnicity, and gender. That is, $Y_{i1}(Z_{i0} = 1), Y_{i1}(Z_{i0} = 0) \perp X_{i0} \,|\, \boldsymbol{W_i}$, where $\boldsymbol{W_i} =$ [client i's age at intake, intake year, race/ethnicity, and gender]^T.

We use one-to-one matching without replacement or ties, which means that any given unit can only be in one matched pair. If there are multiple possible matches for one unit, one of these matches is arbitrarily chosen. Although this approach results in dropping more observations, it allows us to circumvent the computational challenge of incorporating weights in local randomization estimation.[1] Since there are substantially fewer barely ineligible

[1]Stuart (2010) provides an overview of various matching methods and discusses the pros and cons of

individuals than barely eligible individuals (425 vs. 737) in the cleaned dataset, we will be finding matches for the "control" or ineligible individuals among the "treated" or eligible individuals, rather than the other way around. This means that our matched sample will "look like" the "controls." That is, our estimand of interest is the sample average [intention-to-treat] treatment effect on the control (SATC).

We use the `Match` function from the `R Matching` package (R Core Team, 2020; Sekhon, 2011) to identify a sample that is balanced on age at intake, intake year, race/ethnicity, and gender. We use the default Euclidean distance metric and a caliper to ensure that pairs match exactly on race/ethnicity and gender and within two years of age at intake and of intake year. The particular matches identified by the `Match` function depend on the order of the observations in the data frame and the randomization seed. We arbitrarily set a seed for shuffling these data and for matching and are left with 413 matched pairs. This means that we only dropped twelve of the 425 control observations.[2] One of these dropped clients is white and the other is Asian. The other ten barely ineligible clients dropped from the sample are Black; nine are female and one is male. In the matched set, there remain zero white clients, one matched pair of Asian female clients, zero Asian male clients, 39 matched pairs of Black female clients, and 60 matched pairs of Black male clients. Table 6.1 describes the demographics of the matched dataset and includes the results of difference-in-means randomization inference balance checks on this subset of data. Appendix A contains bar charts of the distributions of these covariates after matching.

---

many-to-one matching and matching with replacement.

[2]Over 1,000 repetitions of the matching process with different seeds, between 9 and 15 control observations were dropped each time. The median number dropped was 12.

Covariate	Barely Eligible ($m_1 = 413$)	Barely Ineligible ($m_0 = 413$)	Difference-in-Means	Randomization P-Value
Age at Intake	12.0	12.0	0.0	0.716
Intake Year	2016.7	2016.7	0.0	0.864
Percent Asian	0.2	0.2	0.0	1.000
Percent Black	24.0	24.0	0.0	1.000
Percent Latinx	73.8	73.8	0.0	1.000
Percent Multi-Racial/Ethnic	1.7	1.7	0.0	1.000
Percent Other Race/Ethnicity	0.2	0.2	0.0	1.000
Percent White	0.0	0.0	0.0	1.000
Percent Female	36.6	36.6	0.0	1.000
Note:				$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 6.1: Pre-Treatment Covariate Balance Checks After Matching using Randomization Inference with 10,000 Permutations of the Eligibility Vector $\boldsymbol{Z_0} = (Z_{(1)0}, Z_{(2)0}, ..., Z_{(M)0})^T$

CHAPTER 7

Results

The outcomes tested are the risk score, the raw scores for each of the nine components of the risk score, a binary indicator of eligibility for secondary prevention, and binary indicators of meeting or exceeding the raw score cutoff for each of the nine scales. Our test statistic in all cases is the difference-in-means. The results are summarized in Figures 7.1 and 7.2 and Tables 7.1 and 7.2. Of the 20 outcomes tested, only three difference-in-means are significant at the 5% level. These are the mean raw score for scale A (Antisocial Tendencies), the mean raw score for scale DE (Impulsive Risk Taking), and the proportion of individuals with a concerning raw score for scale G (Negative Peer Influence). Three additional difference-in-means are significant at the 10% level. These are the mean raw score for Scale H (Peer Delinquency), the proportion of individuals with a concerning score for scale DE (Impulsive Risk Taking), and the proportion of individuals with a concerning raw score for scale H (Peer Delinquency). All six of these estimated differences-in-means are positive, meaning that the barely eligible group has a higher mean raw score or higher proportion of individuals with concerning scores than the barely ineligible group. In fact, of the 20 ITT effect estimates, only four are negative. Our analysis fails to detect an effect of initial assignment to secondary prevention as opposed to primary prevention at the local level for this subset of data for most outcomes tested. We expect results to be quite similar for the effect of actual enrollment since compliance rates are high. In this sample, 92.7% of the barely ineligible individuals enroll in primary prevention and 99.8% of the barely eligible individuals enroll in secondary prevention. We used the function `rdrandinf` from the `R` package `rdlocrand` by Cattaneo et al. (2016) to conduct this analysis.

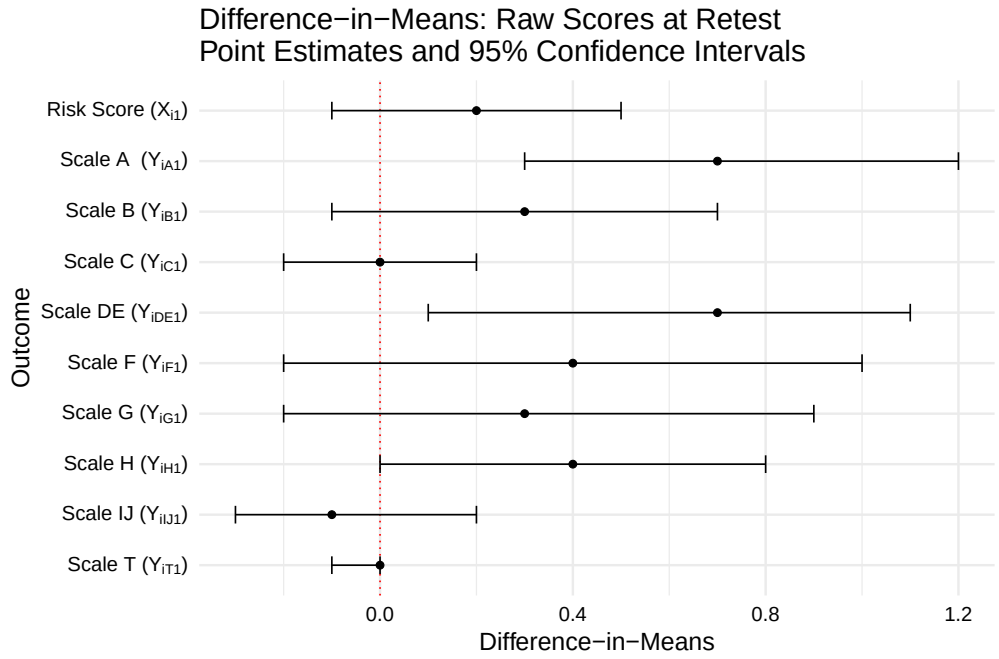

Figure 7.1: Point Estimates and 95% Hypothesis Test Inversion Confidence Intervals for Integer Outcomes Computed from Retest YSETs

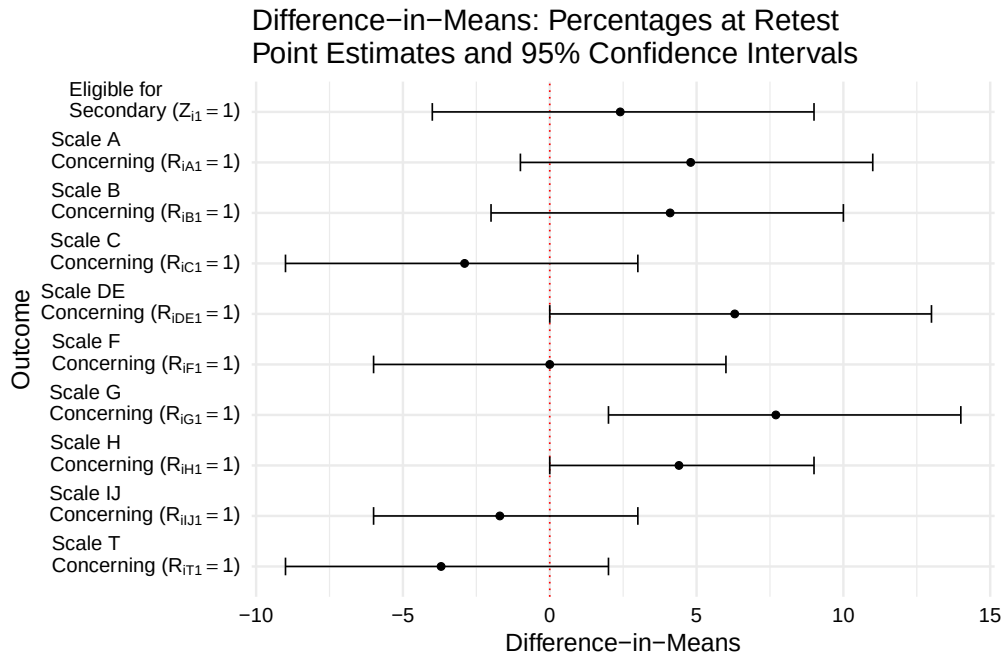

Figure 7.2: Point Estimates and 95% Hypothesis Test Inversion Confidence Intervals for Binary Outcomes Computed from Retest YSETs

Mean at Retest	Barely Eligible ($m_1 = 413$)	Barely Ineligible ($m_0 = 413$)	Difference-in-Means	Randomization P-Value	95% Confidence Interval
Risk Score, X_{i1}	2.9	2.7	0.2	0.243	[-0.1, 0.5]
Scale A Raw Score, Y_{iA1}	14.5	13.8	0.7	0.014**	[0.3, 1.2]
Scale B Raw Score, Y_{iB1}	6.2	5.9	0.3	0.197	[-0.1, 0.7]
Scale C Raw Score, Y_{iC1}	2.7	2.7	0.0	0.863	[-0.2, 0.2]
Scale DE Raw Score, Y_{iDE1}	12.5	11.8	0.7	0.023**	[0.1, 1.1]
Scale F Raw Score, Y_{iF1}	17.0	16.6	0.4	0.202	[-0.2, 1.0]
Scale G Raw Score, Y_{iG1}	11.6	11.3	0.3	0.204	[-0.2, 0.9]
Scale H Raw Score, Y_{iH1}	9.5	9.1	0.4	0.088*	[0.0, 0.8]
Scale IJ Raw Score, Y_{iIJ1}	2.2	2.3	-0.1	0.813	[-0.3, 0.2]
Scale T Raw Score, Y_{iT1}	0.3	0.3	0.0	0.208	[-0.1, 0.0]
Note:					$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 7.1: Difference-in-Means Randomization Inference for Integer Outcomes with 10,000 Permutations of the Eligibility Vector $\boldsymbol{Z_0} = (Z_{(1)0}, Z_{(2)0}, ..., Z_{(M)0})^T$

Percent at Retest	Barely Eligible ($m_1 = 413$)	Barely Ineligible ($m_0 = 413$)	Difference-in-Means	Randomization P-Value	95% Confidence Interval
Eligible for Secondary, $Z_{i1} = 1$	39.2	36.8	2.4	0.509	[-4.0, 9.0]
Score A Concerning, $R_{iA1} = 1$	37.5	32.7	4.8	0.165	[-1.0, 11.0]
Score B Concerning, $R_{iB1} = 1$	39.5	35.4	4.1	0.253	[-2.0, 10.0]
Score C Concerning, $R_{iC1} = 1$	32.7	35.6	-2.9	0.418	[-9.0, 3.0]
Score DE Concerning, $R_{iDE1} = 1$	45.3	39.0	6.3	0.080*	[0.0, 13.0]
Score F Concerning, $R_{iF1} = 1$	37.5	37.5	0.0	1.000	[-6.0, 6.0]
Score G Concerning, $R_{iG1} = 1$	45.5	37.8	7.7	0.024**	[2.0, 14.0]
Score H Concerning, $R_{iH1} = 1$	16.5	12.1	4.4	0.092*	[0.0, 9.0]
Score IJ Concerning, $R_{iIJ1} = 1$	15.5	17.2	-1.7	0.574	[-6.0, 3.0]
Score T Concerning, $R_{iT1} = 1$	23.2	26.9	-3.7	0.264	[-9.0, 2.0]
Note:					$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 7.2: Difference-in-Means Randomization Inference for Binary Outcomes with 10,000 Permutations of the Eligibility Vector $\boldsymbol{Z_0} = (Z_{(1)0}, Z_{(2)0}, ..., Z_{(M)0})^T$

CHAPTER 8

Discussion

In this paper, we used a local randomization regression discontinuity design to estimate and evaluate the average ITT effect of initial assignment to secondary prevention on clients' retest YSET responses. Given the nature of the data and the discrete risk score used to determine eligibility, local randomization RDD is one of the few "quasi-experimental" strategies that can be used evaluate the YSET database, albeit under strong assumptions.

We failed to show a significant ITT effect for 14 of the 20 outcomes tested using this approach. The observed difference-in-means is significant at the 5% level for three outcomes and at the 10% level for an additional three. These six difference-in-means are positive, meaning that the means are more concerning among the barely eligible clients than the barely ineligible clients in our sample. Our analysis relies on strong assumptions and employs finite-sample methods that are only valid at the cutoff, so these results do not rule out the possibility that overall, GRYD Prevention Services may be effective at reducing gang involvement. Further work with this same data includes sensitivity analysis[1] and robustness checks. It may be that our assumption of "as if random" treatment assignment within the window of barely eligible and barely ineligible clients conditional on the four covariates of age at intake, intake year, race/ethnicity, and gender is not adequate. There may be confounding that we are not accounting for, or our results may be biased by the way we chose our sample. A fuzzy RDD could be used for inference on the effects of actual enrollment level, though we do not anticipate that the results would be very different since compliance rates are high.

[1] `rdlocrand` provides functions for computing Rosenbaum confidence intervals under arbitrary interference (Rosenbaum, 2007) and Rosenbaum bounds on sensitivity to unobserved confounders (Rosenbaum, 2002). One can also analyze sensitivity to window length, although we do not believe the "as-if random" assumption is likely to hold for wider windows, and we would also need to address imbalance in the wider windows (Cattaneo et al., 2016).

Nevertheless, is important to continue evaluating GRYD Prevention Services on a regular basis because like medical interventions, well-intentioned social programs can have helpful, harmful, and/or negligible consequences (McCord, 2003; Braga, 2016).

In order to better evaluate GRYD Prevention Services, more data are needed. GRYD is now using Qualtrics to input YSET data, so there should be fewer YSETs with missing scale scores in the future. Although the YSET outcomes are themselves interesting as predictors for future gang involvement, it would be useful to include additional questions in retests. For evaluating the program, unlike for determining eligibility, we do not need to know outcomes for particular individuals, so we can use randomized response techniques (RRT) to gather information about sensitive topics without compromising anonymity. These techniques would hopefully elicit more honest answers, since the interviewer does not know which question the interviewee is asked, but researchers can still estimate the distribution of answers to the sensitive question under certain assumptions (Blair, Imai, & Zhou, 2015). It would also be very valuable to follow up with former GRYD clients and with individuals who completed intakes but did not enroll in a prevention program. Although our analysis yielded few significant results, we believe the null results are equally valuable and hope that this study illustrates the need for further research on this far-reaching program.

Appendix A

Figures

A.1 Intake Risk Score Distribution

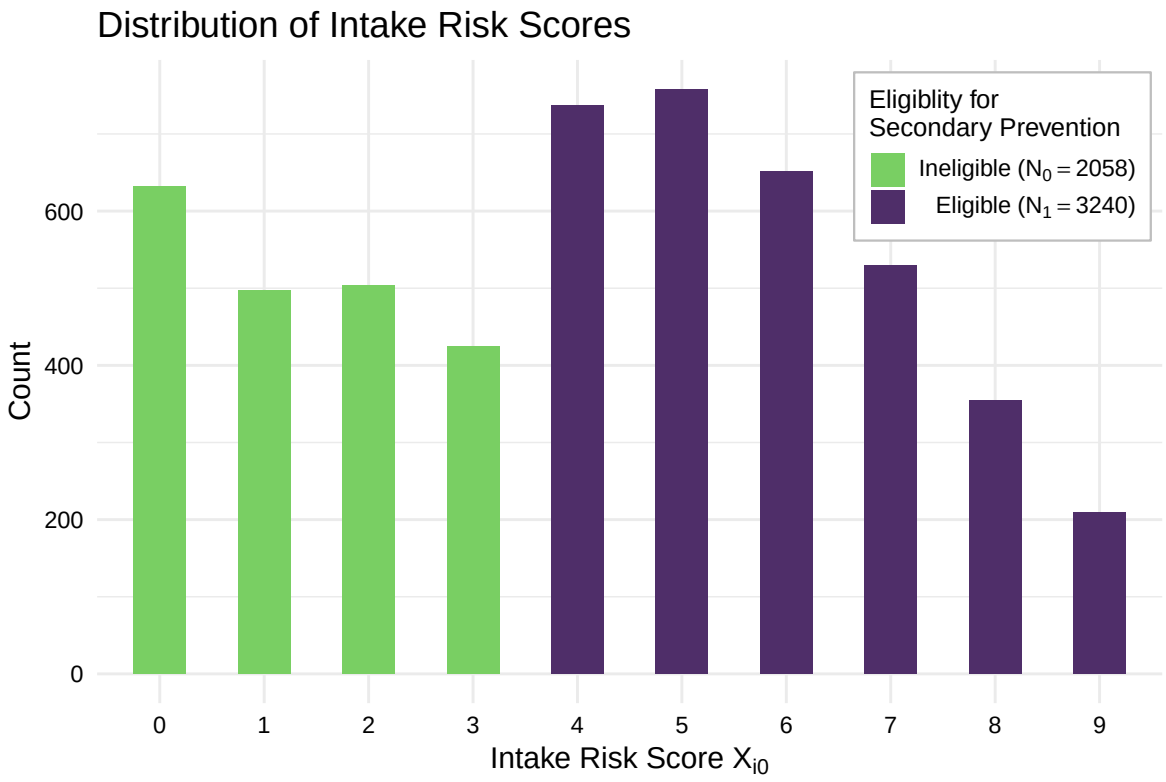

Figure A.1: Distribution of Intake Risk Scores in Cleaned Dataset

A.2 Pre-Treatment Covariate Distributions Before and After Matching

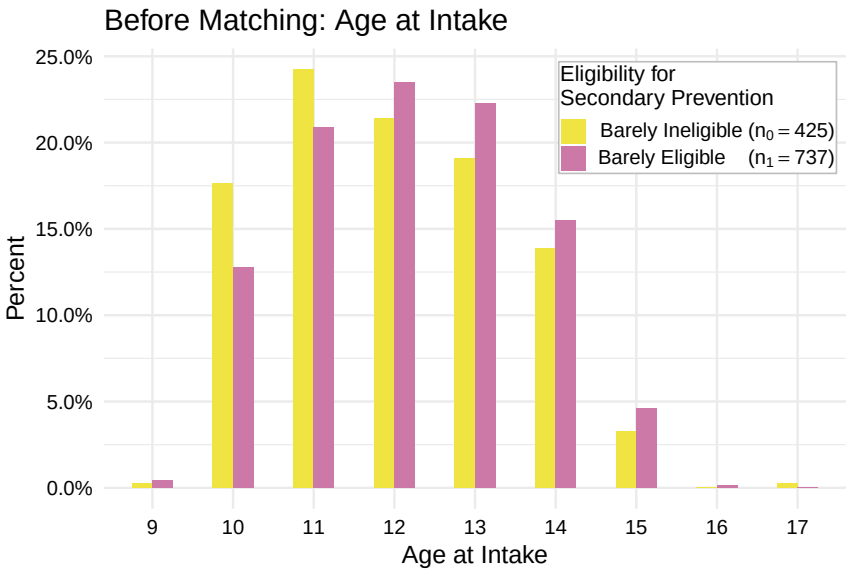

Figure A.2: Age Distribution Before Matching

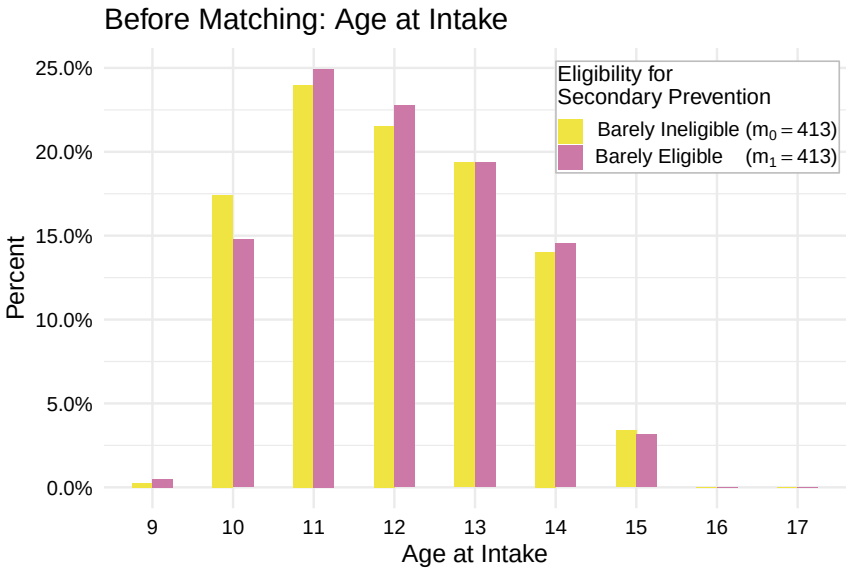

Figure A.3: Age Distribution After Matching

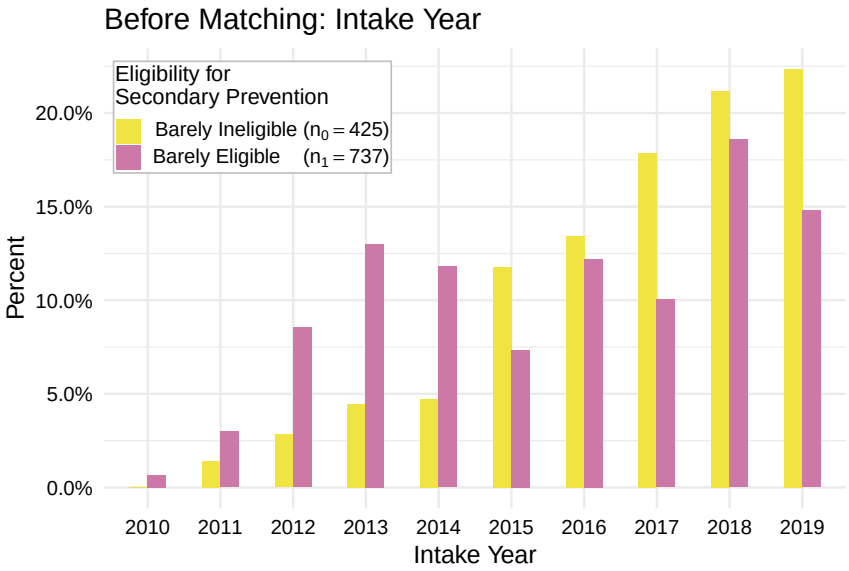

Figure A.4: Intake Year Distribution Before Matching

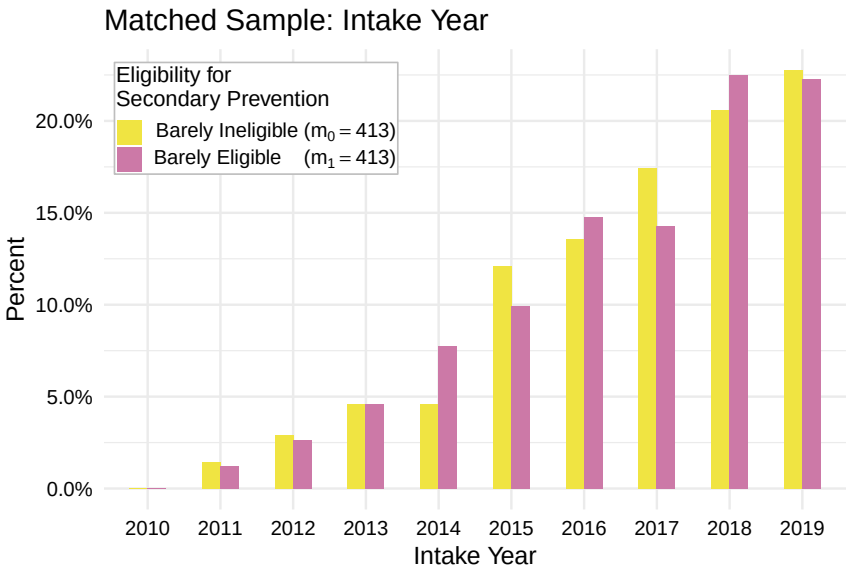

Figure A.5: Intake Year Distribution After Matching

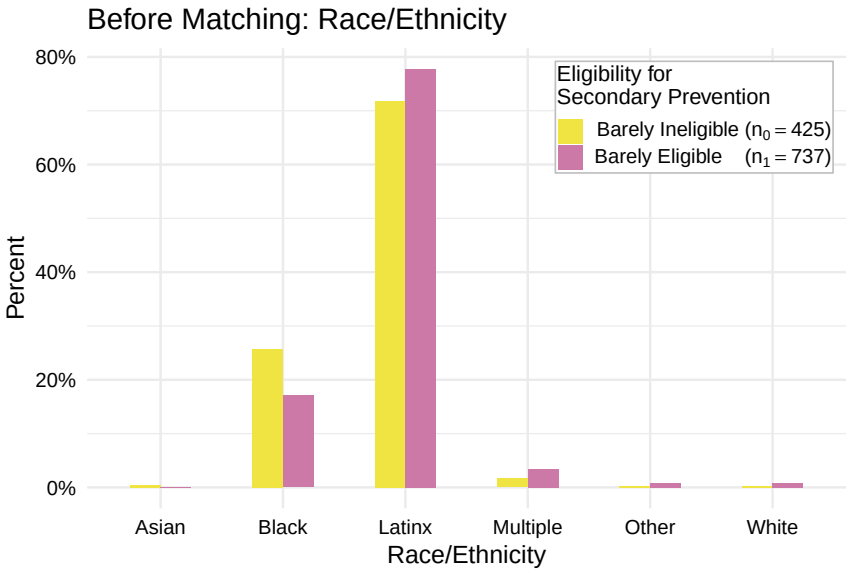

Figure A.6: Race/Ethnicity Distribution Before Matching

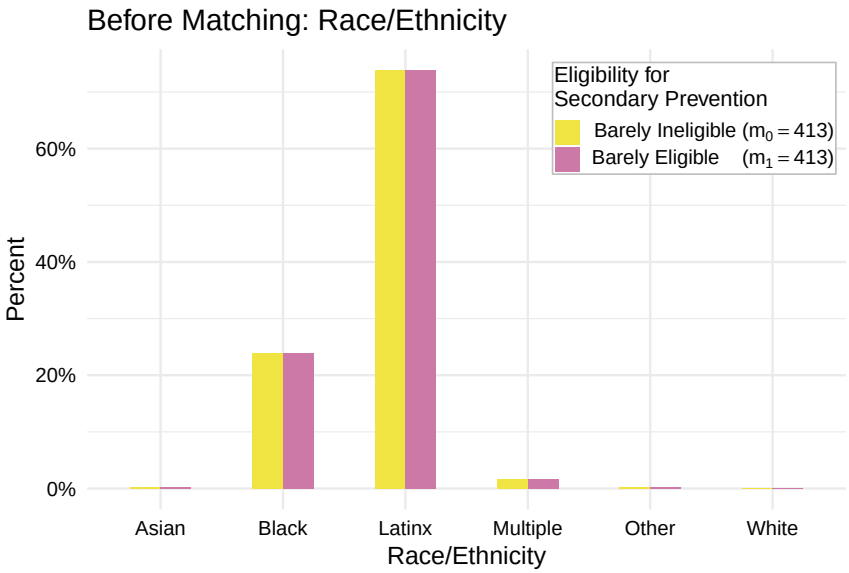

Figure A.7: Race/Ethnicity Distribution After Matching

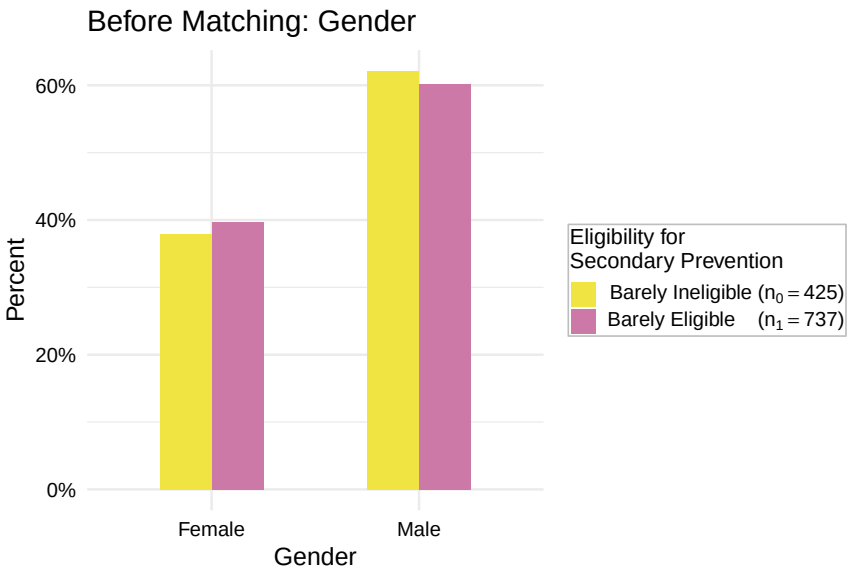

Figure A.8: Gender Distribution Before Matching

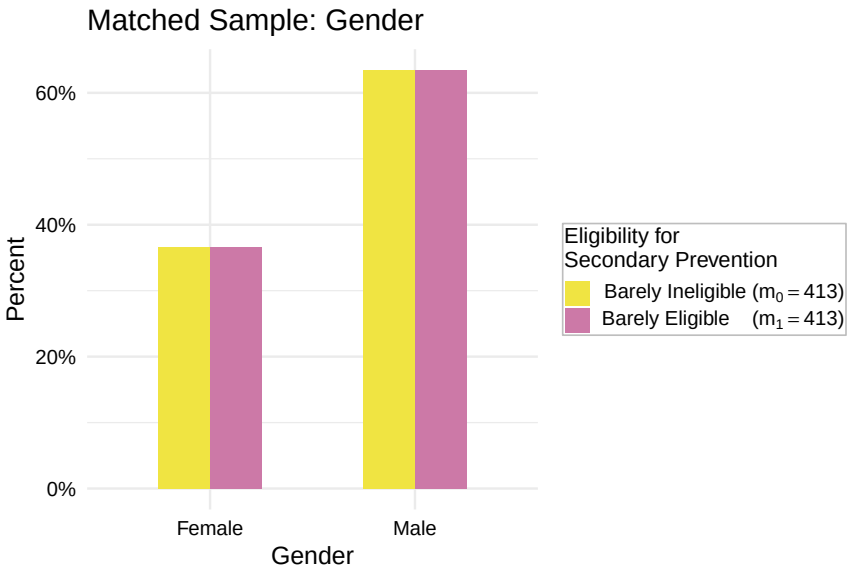

Figure A.9: Gender Distribution After Matching

References

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and Analysis of the Randomized Response Technique. *Journal of the American Statistical Association*, *110*(511), 1304-1319. Retrieved from https://doi.org/10.1080/01621459.2015.1050028

Braga, A. A. (2016). The continued importance of measuring potentially harmful impacts of crime prevention programs: the academy of experimental criminology 2014 Joan McCord lecture. *Journal of Experimental Criminology*, *12*, 1-20. Retrieved from https://doi.org/10.1007/s11292-016-9252-4

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). A Practical Introduction to Regression Discontinuity Designs: Foundations. In R. M. Alvarez & B. N (Eds.), *Elements in Quantitative and Computational Methods for the Social Sciences.* Cambridge University Press. Retrieved from https://doi.org/10.1017/9781108684606

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in Regression Discontinuity Designs under Local Randomization. *The Stata Journal*, *16*(2), 331–367. Retrieved from https://doi.org/10.1177/1536867X1601600205 (Used Version 0.7.1, Published on CRAN on 2020-08-26; R Version 4.0.2)

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2017). Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality. *Journal of Policy Analysis and Management*, *36*(3), 643–681. Retrieved from https://doi.org/10.1002/pam.21985

City of Los Angeles Mayor's Office of Gang Reduction and Youth Development. (2012). *Youth Services Eligibility Tool (YSET).* GRYD Office. (The YSET is the copyright of the City of Los Angeles (2012). All rights reserved. It has been updated over the years, but the overall structure has remained the same.)

City of Los Angeles Mayor's Office of Gang Reduction and Youth Development. (2013, January). *Instructions to Score Youth Services Eligibility Interviews.* GRYD Office.

Cook, T. D. (2008). "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics*, *142*(2),

636–654. Retrieved from https://doi.org/10.1016/j.jeconom.2007.05.002

Dunworth, T., Hayeslip, D., Lowry, S., Kim, K., Kotonias, C., & Pacifici, L. (2013). Evaluation of the Los Angeles Gang Reduction and Youth Development Program: Year 3 Final Report. *Urban Institute and Harder+Company Community Research*. Retrieved from https://www.urban.org/sites/default/files/publication/77951/2000621-Evaluation-of-the-Los-Angeles-Gang-Reduction-and-Youth-Development-Program-Year-3-Final-Report.pdf

Hennigan, K. M., Kolnick, K. A., Vindel, F., & Maxson, C. L. (2015). Targeting youth at risk for gang involvement: Validation of a gang risk assessment to support individualized secondary prevention. *Children and Youth Services Review*, *56*, 86-96. Retrieved from https://doi.org/10.1016/j.childyouth.2015.07.002

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199-236. Retrieved from https://doi.org/10.1093/pan/mpl013

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. Retrieved from https://doi.org/10.1017/CBO9781139025751

Keele, L., Titiunik, R., & Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 223–239. Retrieved from https://doi.org/10.1111/rssa.12056

Kraus, M., Chan, K., Martin, A., Park, L., Leap, J., Rivas, L., ... Kolnick, K. A. (2017). GRYD Gang Prevention 2017 Evaluation Report. *The City of Los Angeles Mayor's Office of Gang Reduction and Youth Development (GRYD) Research and Evaluation Team*. Retrieved from https://www.lagryd.org/sites/default/files/reports/GRYD%20Prevention%20Report_Final.pdf

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, *142*(2), 675–697. Retrieved from https://doi.org/10.1016/j.jeconom.2007.05.004

McCord, J. (2003). Cures That Harm: Unanticipated Outcomes of Crime Prevention Programs. *The Annals of the American Academy of Political and Social Science*, *587*(1), 16-30. Retrieved from https://doi.org/10.1177/0002716202250781

R Core Team. (2020). R: A language and environment for statistical computing, Version 4.0.2 (2020-06-22). *R Foundation for Statistical Computing*. Retrieved from http://www.R-project.org/

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). Springer. Retrieved from https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, *102*(477), 191–200. Retrieved from https://doi.org/10.1198/016214506000001112

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701. Retrieved from https://doi.org/10.1037/h0037350

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*(371), 591–593. Retrieved from https://doi.org/10.2307/2287653

Rubin, D. B. (1986). Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association*, *81*(396), 961-962. Retrieved from https://doi.org/10.1080/01621459.1986.10478355

Sekhon, J. S. (2009). The Neyman-Rubin Model of Causal Inference and Estimation Via Matching Methods. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford Handbook of Political Methodology.* Oxford University Press. Retrieved from https://doi.org/10.1093/oxfordhb/9780199286546.003.0011

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, *42*(7), 1-52. Retrieved from https://doi.org/10.18637/jss.v042.i07 (Used Version 4.9-7, Build Date 2020-02-05; R Version 4.0.2)

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look

Forward. *Statistical Science*, *25*(1), 1–21. Retrieved from https://doi.org/10.1214/09-STS313

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. Retrieved from https://doi.org/10.1037/h0044319