Appendix J: Recidivism Analysis

In quasi-experimental research studies, a comparison group is often used to assess what would have happened had a treatment not been implemented. This is often done retrospectively in social program evaluations. Researchers at JohnJayREC identified a comparison group from New York State criminal history records involving individuals as similar as possible to the BOGAP treatment group. Differences in recidivism (rearrest) were estimated by drawing on a pool of anonymized "top charge" criminal history records maintained by the Department of Criminal Justice Services (DCJS) for males born between 1992 and 2006 who were arrested in 2017 or later for any variation of Penal Law category 265 (criminal possession, use, and/or sale of a firearm and other dangerous weapons).[1] The file contained the universe of individuals with arrest events meeting the criteria, along with a complete set of their previous criminal history records. The initial file included 37,089 reccords organized by each individual's New York State Identification Number (NYSID).[2]

Most records were removed after a data cleaning process found them to be inappropriate as possible comparison cases for various reasons. For example, from the outset of the study, the Bronx District Attorney's Office (Bronx DA) tracked all persons considered for BOGAP. Using the Bronx DA data file, the research team identified 490 persons who had been considered for BOGAP but were not referred for treatment. Those records were excluded from the pool of possible comparison units. Other records were removed for similar reasons. Not all such factors are identifiable, of course. But, assembling a sufficiently large matched comparison group with no known compromising factors mitigates the problem to some extent.

After dropping all records with known factors making them inappropriate as comparison cases, the final comparison pool contained 5,428 units.

---

1 Participants in the BOGAP (treatment) sample were arrested between 2021 and 2024. Restricting comparison cases to arrests occurring between 2017 and 2024 ensures that comparison units are drawn from a time window similar to that of BOGAP participants.

2 NYSID is a unique number, based on fingerprints, assigned by the New York State Division of Criminal Justice Services (DCJS) to a person with justice involvement (criminal and/or civil) in New York State.

Table J.1: Additional Steps of the Data Cleaning Process

Beginning NYSID Pool: 37,089 records 25,230 persons		
Filtering Steps	**Dropped**	**Remaining**
NYSID associated with more than one ID value. If not possible to determine multiple NYSIDs belonged to the same person, records dropped from the analysis.	1,292	23,938
Record not related to Criminal Possession of a Weapon (CPW). DCJS data file provides binary flag for CPW. Non-CPW cases dropped.	217	23,721
Disposition date missing. Cases with missing or invalid disposition dates dropped from sample. Recidivism for comparison cases tracked from disposition.	232	23,489
Arrest level is misdemeanor.	7,133	16,356
Sex of individual is female. BOGAP sample all male.	20	16,336
Individual under age 17. BOGAP sample at least 17.	262	16,074
Record indicates indeterminate incarceration sentence. Release date required to track recidivism.	3,434	12,640
Arrest not between 2017 and 2020, most effective time period for comparison cases similar to those referred to BOGAP.	7,212	5,428
TOTAL UNITS DROPPED	19,802	

Matching Analysis

Coarsened exact matching was used to construct a comparison group similar to the treatment group based on age at arrest and prior arrests. CEM works by grouping important characteristics, such as age at arrest and number of prior arrests, into broader categories. The method then compares individuals only within those groups. If a group contains both program participants and non-participants with similar characteristics, those cases are retained. This is known as common support. Groups that contain only comparison cases are removed because they lack a comparable program participant.

The BOGAP study sample included 84 treated units, and the CEM analysis yielded 4,987 matched control units based on age at arrest and prior arrests. The unmatched units (441) were off support (i.e., these subjects fell outside the range of characteristics observed in the treated group). To illustrate, one stratum contained two treated units matched to 44 control units. And that stratum comprised treated and control subjects between 17 and 19 years of age at the time of arrest, who were charged with a level E felony offense.

Table J.2: CEM Strata, Descriptives

			Size of Strata (Number of Persons)	
Strata	**Arrest Age**	**Arrest Class**	**Treated**	**Matched Comparison**
1	17-20	E Felony	2	124
3	17-20	D Felony	2	794
4	17-20	C Felony	28	1,198
6	21-24	E Felony	1	90
7	21-24	B Felony	1	19
8	21-24	D Felony	1	933
9	21-24	C Felony	39	1,331
13	25-27	C Felony	10	498
		TOTAL	**84**	**4,987**

Each comparison case is assigned a weight to ensure the comparison group reflects the distribution of characteristics observed among program participants. The matched sample is then used to estimate differences in recidivism outcomes between program participants and the comparison group. Within each matched strata, treated and control units are then comparable on coarsened covariates – arrest class (severity of the arrest charge) and age at arrest. CEM weights controls so each stratum contributes proportionally to the treated group. CEM does not compare individuals directly. It compares treated versus weighted controls within strata, then aggregates across strata, so large control pools (N=4,987) do not dominate the analysis. Everything is scaled to match the treated distribution. Each stratum comprises a number of control units (persons) that together constitute a single comparable unit; hence, the need to statistically weight their contributions. CEM essentially creates a synthetic sample where treated and control units are comparable.

Before matching, the arrest age and arrest class of the treated group differed significantly from the average value in the pool of controls, and variances in the unmatched pool were less than half that of the treated group. After matching, the mean difference in arrest age and class was essentially eliminated. The bins created during CEM enable exact alignment of values. The L1 statistic is a global imbalance measure based on differences between treated and control distributions over the multivariate histogram of matching covariates (arrest age and arrest charge severity in this study). Lower L1 values mean better balance, and a reduction from pre- to post-matching indicates improved balance. Before matching, L1 was 0.413; after matching, it was reduced to 0.143, indicating that CEM substantially improved overall multivariate balance.

Table J.3: CEM Balance Measures

	BEFORE MATCHING		AFTER MATCHING	
	Imbalance	**Variance Ratio**	**Imbalance**	**Variance Ratio**
Arrest Age	–0.0674 (large imbalance)	0.7969	–0.0057 (excellent balance)	0.9718 (almost identical)
Arrest Class	0.8776 (large imbalance)	0.5193	–0.0000 (perfect balance)	1.0117 (perfect)

- Balance measures before CEM (matching) compare the treated group to the unmatched pool of comparison units, and, after CEM, compare the treated to the matched control pool.
- Imbalance indicates the mean difference of a given covariate. Values close to 0 are preferred.
- Variance Ratio values close to 1 are preferred after CEM and indicate a very similar distribution of values of a given covariate between the treated and matched control units.
- Overall Multivariate Imbalance Measure: L1=**0.413** (before matching)
- Overall Multivariate Imbalance Measure: L1=**0.143** (after matching)
- Percentage of local common support: LCS=**58.1%**

Even though CEM matching identified 4,987 control units, after weighting they behave like ~3,359 independent observations. Raw counts do not constitute the properly weighted post-CEM comparison group. For any post-matching estimation of an effect, controls should be weighted so they match the distribution of treated units across strata. The weights rebalance the contribution of control units in large strata. Without weights, controls would dominate large strata, thereby skewing results. Table J.4. describes characteristics of the CEM-derived matched comparison group. Optimal balance between BOGAP-treated and comparison groups was achieved by matching on arrest age and arrest class. Researchers subsequently controlled for region and criminal history (total prior misdemeanor-level arrests) in the Cox Proportional Hazard model, ensuring that variables not included in the matching process were accounted for in the recidivism analysis.

Table J.4: Sample Characteristics and Prior Criminal History

	Treated (N=84)		Matched Comparison (N=4987)	
	Number	Percent	Number	Percent
Age at Arrest (BOGAP Arrest)				
17	2	2%	302	6%
18-20	30	36%	1598	32%
21-23	35	42%	1962	39%
24-27	17	20%	1125	23%
Arrest Class (BOGAP Arrest)				
C Felony	77	92%	4571	92%
D Felony	3	4%	178	4%
E Felony	3	4%	178	4%
B felony	1	1%	59	1%
Arrest Region (BOGAP Arrest)				
NYC	84	100%	4756	95%
Rest of State	0	0%	231	5%
Prior Misdemeanor Arrests				
0	49	58%	2435	49%
1-2	26	31%	1152	23%
3+	9	11%	1399	28%

Note: For the CEM comparison group, weighted counts are presented and may not always sum exactly to the total N of 4,987 due to rounding.

Multivariate Modeling

To examine factors associated with the likelihood of rearrest, researchers estimated a Cox proportional hazards model. This type of model is appropriate when the outcome of interest is the time until an event occurs, rather than simply whether the event occurs. The model included the following variables:

- Treatment - Indicates whether an individual participated in BOGAP
- Number of misdemeanor arrests - A measure of prior criminal history
- Age at arrest - Known predictor of recidivism
- Region indicator - Controls for geographic differences in arrest patterns. NYC vs. Rest of NY.

Results are presented as hazard ratios (exponentiated coefficients). Values greater than one indicate higher risk of rearrest. Values less than 1 indicate lower risk, and values equal to 1 indicate no effect. The findings indicate that participation in BOGAP was not associated with a statistically significant difference in rearrest risk compared with the comparison group. The strongest predictor of rearrest was prior criminal history (12% increase), followed by age at arrest (7% decrease). The research team attempted to improve overall statistical power by using more than one matched comparison unit per treated person and weighting the larger comparison pool in subsequent analyses. The main limitation is that the treated group experienced only 23 events during the observed period. The count of events (rearrests) in the treated group can be thought of as the ceiling on the precision the study design can realistically achieve.

Table J.5: Cox Proportional Hazards Model (Time to Rearrest)

Sample = 5,071 Number of Events= 2,870 arrests (23 treated, 2,847 matched comparison)

	Hazard Ratio	Risk	Lower 95% CI	Upper 95% CI	Standard Error (SE)	Robust SE	z statistic	p-value
Treatment (BOGAP)	0.94	-6%	0.62	1.41	0.210	0.209	-0.31	0.757
Prior Misdemeanor Arrests	1.12	12%	1.09	1.15	0.004	0.015	7.808	<0.001
Age at Arrest	0.93	-7%	0.88	0.98	0.019	0.026	-2.938	0.003
Rest of the State (vs. NYC)	0.96	-4%	0.76	1.21	0.100	0.119	-0.334	0.738

Model Diagnostics:
Concordance = 0.68 (se = 0.007) / Likelihood ratio test = 580.6 on 4 df, p=<2e-16 / Wald test = 63.55 on 4 df, p=5e-13 / Score (logrank) test = 975.2 on 4 df, p=<2e-16, Robust = 446.6 p=<2e-16.

Notes: The likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not. Hazard ratios are the exponentiated form of the coefficients estimated or generated by the model.