Preliminary Report: Racial and Socioeconomic Factors in Predictive Policing

EAINT MAY, NORA PHELAN, ZACK TAYLOR, ASH TODD, CMPU 250, Vassar College

1 ABSTRACT

Intimate partner violence is a complex issue which is difficult to predict due to the many factors that may influence individuals to commit it. In this study we aim to enhance the accuracy of predictive policing on this kind of violence by analyzing NYPD domestic violence reports. We take a place-based approach which focuses more on policing by area and therefore mitigates broader biases involved with race, gender, and other personal characteristics. We find that there are many overlaps in this place-based approach where racial and economic factors are disproportionate in areas of high police activity. Our findings present many obstacles for the deployment of fair and accurate predictive policing measures. Systemic issues like racial bias and socioeconomic disparity need to be addressed before a truly fair system of predictive policing can be created.

2 KEYWORDS

Predictive Policing. According to Brennan Center for Justice: "Predictive policing involves using algorithms to analyze massive amounts of information in order to predict and help prevent potential future crimes." (Diaz, 2019)

DIR. Short for Domestic Incident Report, the report required to be filled out by the NYPD when investigating a domestic issue.

IPV. Short for Intimate Partner Violence, defined by the WHO as "behaviour within an intimate relationship that causes physical, sexual or psychological harm, including acts of physical aggression, sexual coercion, psychological abuse and controlling behaviours."

Intersectionality. The Merriam-Webster Dictionary defines intersectionality as "the complex, cumulative way in which the effects of multiple forms of discrimination (such as racism, sexism, and classism) combine, overlap, or intersect especially in the experiences of marginalized individuals or groups."

3 INTRODUCTION

Domestic violence is a deeply complex issue, rooted in histories of inequality, shaped by intersecting forces of gender, economic status, and systemic neglect. Domestic violence is officially defined as "physical, sexual, emotional, economic, psychological, or technological actions or threats of actions or other patterns of coercive behavior that influence another person within an intimate partner relationship." (OVW). The variety of social and psychological factors which can push someone to commit violence within the home make it difficult to predict future offenses. As a result, responses to domestic violence are often lackluster and do not do enough to prevent the issue from recurring. In many cases, this failure to produce concrete assistance incentivises other victims not to report violence, and therefore makes fixing the broader issues even tougher.

This project aims to decipher the ways in which police data may be used to allocate resources to combat intimate partner violence. Through the analysis of the New York City Police Department's official reports of domestic violence, we were able to create a model of a place-based predictive algorithm pertaining to IPV reports. To achieve this we began by detailing the steps that a city such as New York might take to develop a predictive model for IPV: what data would be collected, how it would be processed, and what features might be prioritized. Using publicly available police report data, we explore the patterns and correlations that might inform such a model. We will also interrogate the assumptions behind this data and ask what gets lost, distorted, or reinforced when IPV is approached through a predictive lens. What does it mean to make violence "predictable" using data that itself is shaped by uneven policing and systemic bias? And who bears the risk when predictive tools are used in already-vulnerable communities?

Unfortunately, as many cases go unreported for a variety of reasons, our research is constrained to reported cases of domestic violence. In the case of intmate partner violence it is estimated that "each year, approximately 500,000 women are physically assaulted or raped by an intimate partner compared to 100,000 men." (National Library of Medicine, 2023).

4 DATA

The dataset used in our analysis was released to the public by the New York City Police Department. It contains data recorded from domestic violence related offenses in New York City from 2020 and 2021. Each report contains information on the type of offense, the date it was reported, the precinct code and borough in which it occurred, if the offense involved an intimate partner, the race, sex, and age of the victim and offender, and information on the financial state of their area. The type of offense is listed as either a Domestic Incident Report (DIR), felony rape, or felony assault. A DIR is required to be filled out every time the NYPD responds to a domestic incident. They are taken at the time of the response and contain no information on further conviction or other legal interventions involving either party (suspect or victim).

Prior to cleaning, the dataset contained missing values in multiple columns. The columns indicating high poverty, low median income, and high unemployment either contained a 1 or were left blank. These NaNs were changed to 0. Additionally, the columns containing the reported ages of the suspect and the victim had a significant number of blank entries. This could be due to insufficient information at the scene of the incident or mistakes in the reporting process. Regardless, we felt that this missing information could cause inaccuracies in our analysis when it came to studying overall themes in age difference against other factors. For this reason we took the average age in each column and replaced the missing values with those averages. This way we could take the ages across the entire dataset against other factors and still produce reasonably accurate results. Any remaining rows containing NaNs in other columns, as well as one obviously incorrect age (likely a human error), were removed from the dataset.

The datatypes of multiple columns also needed to be changed in order to be used in our analysis. The suspect and victim age columns were changed from strings to integers, with 0 indicating female and 1 indicating male. Suspect and victim ages were also changed to integers, and the report date column was changed to datetime.

For the purposes of our analysis, we needed a way to determine if the reported offense was a felony or not. The existing offense type column has three possible values: DIR, Rape, or Felony Assault. The latter two constitute felonies, but a DIR can result in a misdemeanor, a felony, or no arrest. To simplify this data, we created a new variable called Felony Offense. This variable attaches a boolean value to the report: 1 for a reported felony (either rape or felony assault) or 0 for no reported felony (DIR). This way we can easily make connections between a reported crime and other variables within the data.

5 METHODS

According to a report on New York City Surveillance Technology by the Brenna Center, there are two types of predictive policing in place: place-based and person-based. Place-based predictive policing uses algorithmic systems to analyze datasets to try to predict where certain crimes are likely to occur. Police presence is then deployed based

Precinct	Reports	High Felonies
46	293	True
75	252	True
40	246	True
43	237	True
47	235	True
73	225	False
44	203	True
48	186	False
52	185	True
67	183	True

Table 1. Precincts with Highest Reports in February 2020

on the predictions. Person-based predictive policing uses algorithmic systems to analyze datasets to generate a list of individuals that are likely to commit a crime. (Diaz, 2019)

A model of place-based predictive policing was created by finding the ten precincts with the highest number of DIRs for a given month. A precinct appearing on this list may indicate a place with a high number of domestic incidents. The ten precincts with the highest number of felonies for the given month were also tracked. These two lists of precinct codes were compared and an additional flag was placed next to high-reporting precincts that also contained a high number of felony reports. This information could be used by the New York City Police Department to predict "hotspot" locations. The department's response would then be to allocate additional resources to these locations and increase patrol units in the area. (Developing the NYPD's Information Technology)

Another area of our analysis was the racial makeup of the suspects and victims from each precinct. We created a chart to show the racial proportions of the victims from each precinct, and then a similar chart showing the racial proportions of the suspects. Displaying this information alongside the volume of reports in each precinct can help to determine a connection between predictive policing and racial bias. It could be used in these "hotspot" precincts to determine if there is a racial bias in the policing of these areas or, when compared to the area's actual population, if the racial makeup of those precincts is driving that assumption.

Finally, the reports in the dataset were examined on the basis socioeconomic standing. We calculated the percentage of community districts in each borough that are economically disadvantaged, including the percent of community districts experiencing high poverty, high unemployment, and a low median income. We also calculated the percentage of the total number of reports in each borough that occurred in an economically disadvantaged community district, again based on poverty, unemployment, and median income. This information allows us to draw connections between the rate of policing and economic inequality.

6 RESULTS

In the created model, a total of 6 months were analyzed–February through April in both 2020 and 2021–and 8 precincts were consistently present in the resulting reports: precincts 75, 43, 47, 40, 46, 73, 52, 67. Precincts 75, 73 and 67 are located in Brooklyn and 43, 47, 40, 46 and 52 are located in the Bronx. All of these precincts were also found to have a high felony rate for at least one out of the three months analyzed for each year.

Based on our bar graphs detailing the racial makeup of the reports from each precinct, we did not observe any significant difference between victims and suspects. Most of the precincts showed a black majority, although a few

Precinct	Reports	High Felonies
46	324	True
75	285	False
73	283	False
43	259	True
47	238	True
40	227	True
42	225	True
44	203	True
52	201	True
67	190	False

Table 2. Precincts with Highest Reports in March 2020

Precinct	Reports	High Felonies
75	261	False
46	231	True
73	215	True
43	213	True
40	206	True
42	205	False
44	201	True
48	185	False
47	182	False
52	158	False

Table 3. Precincts with Highest Reports in April 2020

Precinct	Reports	High Felonies
75	272	True
46	271	True
43	218	False
73	210	Flase
67	210	True
47	207	True
44	199	True
42	189	True
40	181	False
52	178	False

Table 4. Precincts with Highest Reports in Feb 2021

had a strong white majority. This remained true across both victims and suspects for each precinct. Additionally we observed one percent with a majority–both victims and suspects–reported as Asian/Pacific Islander: precinct 106.

Our socioeconomic data shows that the Bronx, Brooklyn and Manhattan had the highest percentages of community districts with a high poverty rate and a low median income, while the Bronx, Manhattan and Queens had the highest percentages with high unemployment. These rankings were similar to the boroughs with the highest percent of reports

Precinct	Reports	High Felonies
75	299	False
46	294	True
73	283	True
43	240	True
47	234	True
44	229	True
42	228	True
40	218	True
67	189	False
52	187	False

Table 5. Precincts with Highest Reports in March 2021

Precinct	Reports	High Felonies
46	333	False
75	297	True
73	264	True
43	249	True
40	237	True
67	196	False
47	196	True
42	192	False
52	190	True
105	188	False

Table 6. Precincts with Highest Reports in April 2021

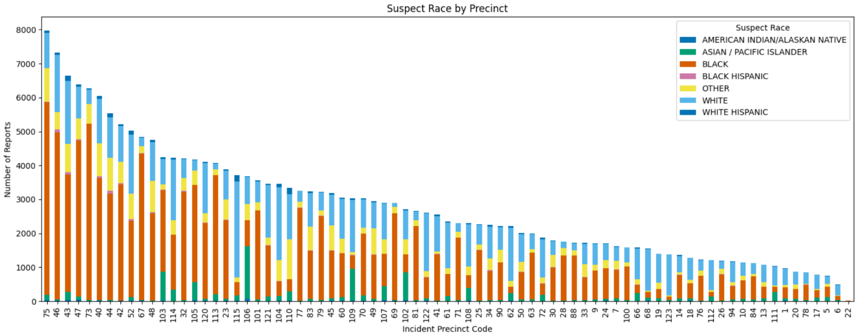

Fig. 1. Bar graph of racial makeup of suspects

occuring in economically disadvantaged community districts, though their order did vary, and the percentage of reports was generally higher than the percentage of districts.

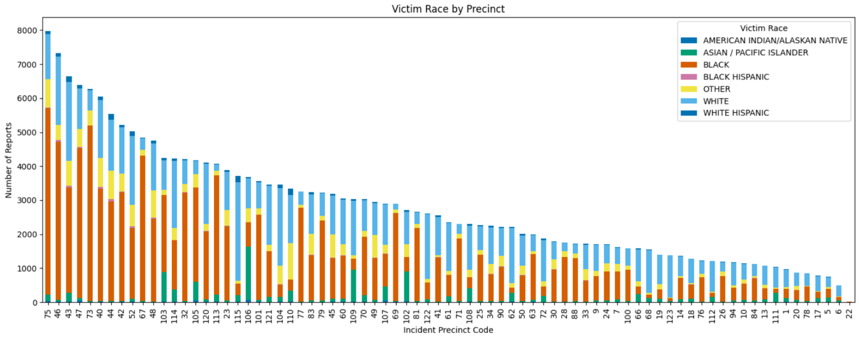

Fig. 2. Bar graph of racial makeup of victims

	Poverty	Unemployment	Median Income
Bronx	0.747846	0.909770	0.747846
Manhattan	0.466064	0.470966	0.466064
Brooklyn	0.397933	0.099839	0.446663
Queens	0.000629	0.167266	0.000629
Staten Island	0.000000	0.000000	0.000000

Table 7. Percent of Reports Occurring in Economically Disadvantaged Community Districts

	Poverty	Unemployment	Median Income
Bronx	0.533333	0.666667	0.533333
Brooklyn	0.250000	0.050000	0.300000
Manhattan	0.230769	0.307692	0.230769
Queens	0.000000	0.055556	0.000000
Staten Island	0.000000	0.000000	0.000000

Table 8. Percentage of Community Districts at an Economic Disadvantage

7 DISCUSSION

The consistency of high report numbers across specific precincts points to two boroughs of New York City: Brooklyn and The Bronx. These figures suggest that a predictive policing system based on similar data would disproportionately target these areas, reinforcing or creating a reputation for criminality. This, in turn, could lead to heightened suspicion of the residents. Notably, these boroughs also tend to experience higher levels of poverty and unemployment.

Additionally, the higher percentage of reports from economically disadvantaged community districts, as compared to the percentage of districts at a disadvantage, could be an indicator of systemic bias. If each community district had a relatively similar level of reports, the percentage of reports from disadvantaged districts would be the same as the percentage of districts at a disadvantage. The fact that the percentage of reports is higher indicates a higher police presence in poorer districts.

Our analysis of the data shows systemic issues related to the racial and socioeconomic makeup of New York City which could cause bias in predictive policing. This presents ethical issues with any attempt to provide a system of predictive policing as it may inflate these systemic problems. Attempts to increase police presence in an area based on a model trained on systemically unfair data will lead to more reports in that area, and therefore create more unfair data to inform said model.

Understanding how intersectionality affects domestic violence is critical to properly address the issue in an effective manner. Factors such as race, gender, class, and welfare status shape lived experiences, struggles, and access to support systems, and therefore useful methods of intervention (Josephson, 2002). By understanding the impact of intersectionality, we can further identify the disparities in reporting, response, and prevention strategies and develop domestic violence interventions that address the needs of all affected communities.

Using entirely government-collected data presents limitations to our analysis. High representation of Black individuals in both victim and suspect roles mirrors historic patterns of over policing in predominantly Black neighborhoods. A model trained on this data may overfit to racially coded patterns of surveillance, reinforcing the notion that domestic violence is more likely to occur in Black communities- not because of actual prevalence, but because of systemic inequities in reporting and enforcement. This must be taken into consideration when presenting any findings from analysis on data such as this.

The absence of population data, specifically the populations of different community districts, populations within the jurisdiction of precincts, and the racial makeups of these populations, limits the conclusions we can draw from our data. Differences in the number of arrests in poorer districts could be due to over policing, but they could also be the result of differences in population. Incorporating this population data could help to rule out concerns of inaccurate findings based on differences in the distribution of races in the areas being studied.

8 REFERENCES

2020 report on the intersection of domestic violence, ..., January 13, 2023. https://www.nyc.gov/assets/ocdv/downloads/pdf/endgbv-intersection-report.pdf.

“Developing the NYPD’s Information Technology.” NYC Government . Accessed April 12, 2025. https://www.nyc.gov/html/nypd/html/home

Díaz , Angel. “New York City Police Department Surveillance Technology.” Brennan Center for Justice, October 4, 2019. https://www.brennancenter.org/our-work/research-reports/new-york-city-police-department-surveillance-technology.

“Domestic Violence.” Office on Violence Against Women (OVW), January 22, 2025. https://www.justice.gov/ovw/domestic-violence.

Francis, Chloe. Et. All, “Algorithmic Accountability: The Need for a New ...” Yale, January 18, 2022.

Josephson, Jyl. “The Intersectionality of Domestic Violence and Welfare in the Lives of Poor Women.” Journal of Poverty 6, no. 1 (January 2002): 1–20. https://doi.org/10.1300/j134v06n01$_0$1.

Lipsky, Sherry, Raul Caetano, and Peter Roy-Byrne. “Racial and Ethnic Disparities in Police-Reported Intimate Partner Violence and Risk of Hospitalization among Women.” Women’s Health Issues 19, no. 2 (March 2009): 109–18. https://doi.org/10.1016/j.whi.2008.09.005.

Monterrosa, Allison E. “How Race and Gender Stereotypes Influence Help-Seeking for Intimate Partner Violence.” Journal of Interpersonal Violence 36, no. 17–18 (June 13, 2019). https://doi.org/10.1177/0886260519853403.

New York State, Capital View Office Park. “NCJRS Virtual Library.” Domestic Incident Policy: Model Law Enforcement Policy Language | Office of Justice Programs, September 2023.

Sokoloff, Natalie J., and Ida Dupont. “Domestic Violence at the Intersections of Race, Class, and Gender.” Violence Against Women 11, no. 1 (January 2005): 38–64. https://doi.org/10.1177/1077801204271476.