WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Fact Sheet Series by Professor Michael T. Light

Racial Disparities in California Criminal History Data
No.000: Explanation of the California CORI Data[1]

INTRODUCTION

This general factsheet explains the use of California Criminal Offender Record Information (CORI data) by the University of Wisconsin (UW) at Madison Department of Sociology to examine racial disparities associated with particular charges in California over a 13-year period (2006-2018). This factsheet should be read alongside the related individual factsheets.

SCOPE AND SOURCE OF DATA

Scope of CORI Data. CORI data are constituted of criminal histories of individuals who come into contact with law enforcement in the state of California. They are equivalent to the aggregation of the records of arrests and prosecutions (or "RAP sheet" information) gathered from law enforcement agencies in California. RAP sheet information includes socio-demographic characteristics of arrested individuals, their criminal histories, state statutes underlying any arrest, the originating agency of any arrest, and case dispositions (or final outcomes), including convictions, dismissals, and sentences.

Source of CORI Data. CORI data are compiled by law enforcement agencies and collected and maintained by the California Department of Justice (CalDOJ). UW-Madison has access to CORI data from 2006 to 2018 through a research agreement entered into with the CalDOJ Justice Information Services Division Research Center.[2]

Primary Scope of Analysis. UW-Madison obtained all CORI data from 2006 to 2018, including all California criminal histories of individuals with law enforcement interaction during this time, as well as their arrest and conviction records pre-dating 2006. We selected felony and misdemeanor arrests of adult defendants (defined as aged 18 or above) and excluded all arrests that only concerned violations of federal immigration law from our analysis.

Cleaning the Data. For approximately 35% of the arrests included in the CORI dataset, there was no subsequent history or disposition outcome. As a default cleaning method, we treated these as law enforcement releases or case dismissals by the prosecutor. For all cases where there is prosecutorial data indicating an initial charge in the dataset, the missing data regarding subsequent case history or disposition outcome was relatively minor, constituting less than 5% of the data. (See table below.) We excluded these records as part of the cleaning process. This decision did not substantially influence the main observed characteristics of the cases. The total number of cases after the data cleaning process is 19,112,520 cases. See the descriptive statistics table below for more information.

CASE OUTCOME VARIABLES

To analyze case outcomes in the California criminal legal system, we considered arrests and four important milestones of criminal case processing: 1) case acceptance, 2) conviction, 3) incarceration, and 4) sentence. We also created indicators for other criminal sanctions.

Case Acceptance. We defined case acceptance as whether a prosecutor accepted a case referred by law enforcement authorities and filed charges in court for the violation of a state statute, or instead declined to file a charge against the individual.

Conviction. If the prosecutor proceeded with a criminal charge, the data allowed us to measure if the case resulted in conviction. Cases that ended in dismissal or acquittal were not counted as convictions. We included deferred adjudication as a conviction.

Incarceration. For those cases that resulted in conviction, we measured the resulting criminal sanction, including jail time, prison time, and/or a fine. We defined incarceration to include where the case resulted in an unconditional incarceration in local jail or state prison. One issue here is that in some cases, the disposition was a choice between jail and a fine, and the CORI dataset does not include the choice ultimately made by the defendant between these two options. We conservatively defined incarceration to include only unconditional incarceration, and to exclude such cases where a defendant could choose a fine in lieu of incarceration.

Length of Incarceration. If the individual received a sentence of incarceration, we measured the length of sentence. (We assigned "0" for non-incarcerated individuals.) In line with the guideline of the U.S. Sentencing Commission,[3] we translated life imprisonment (and extremely long sentences such as 120 years of imprisonment or more) to 480 months to avoid potential outliers in the data. In some analyses, for instance in analyzing severe crimes, we may consider separately instances of life imprisonment and extremely long sentences because of their relevance to the research questions.

Other Criminal Sanctions. We analyzed other criminal sanctions, such as deferred adjudication on condition of treatment or other measures, (the length of) probation, and whether the case led to a fine.

NON-DISPOSITION VARIABLES

The CORI dataset includes the following individual variables as provided by law enforcement reporters.

Age. Age in the CORI dataset is measured at the time of arrest for the charge at issue.

Gender. Gender in the CORI dataset is broken down into only three categories: male, female, and unknown.

Race. The original race variables in the CORI dataset included the following categories: Asian Indian, Black, Chinese, Cambodian, Filipino, Guamanian, Hawaiian, Hispanic, Japanese, Korean, Laotian, Native American, other race, other Asian, Pacific Islander, Samoan, Vietnamese, and White. We collapsed some of these original categories -- to Asian (original categories: Asian Indian, Chinese, Cambodian, Filipino, Guamanian, Hawaiian, Japanese, Korean, Laotian, other Asian, Pacific Islander, Samoan, Vietnamese), Black, Hispanic, White, Other Races (original categories: Native American, other race) – to have a sufficient number of cases in each category to allow for analysis. Each racial category is exclusive (i.e., only one category of race is listed for each individual in the dataset). There is another category in the dataset for unknown race.

NON-DISPOSITION VARIABLES (CONT'D)

Place of Birth. The CORI dataset includes place of birth. We collapsed the detailed categories (i.e., with names of countries) into binary variables indicating whether the individual was born in the United States or a foreign country.

Citizenship. The country of citizenship variable in the CORI dataset is provided by law enforcement reporters. We collapsed the detailed categories (i.e., with names of countries) into binary variables of whether the individual was a U.S. citizen or non-citizen. Although over half of the cases were missing this variable, when we classified as U.S. citizens those who were listed as born in the United States, the rate of missing data decreased to under 3%.

Criminal History. We categorized criminal history by three variables: (1) the number of previous felony arrests in California; (2) the number of previous misdemeanor arrests in California; and (3) the number of previous incarcerations in California. To avoid potential outliers in the data, we top-coded each of these variables at 10, meaning that an individual with more than 10 arrests would be recoded as "10."

Case Characteristics. To measure offense characteristics, we used two variables: (1) the number of charges associated with the arrest (top-coded at 6, to avoid potential outliers in the data, meaning that an individual with more than 6 associated charges would be recoded as "6."); and (2) the offense code of the most serious charge (with seriousness defined by the average incarceration length).

Originating County of Arrest and Disposition. The CORI dataset has an indicator for the record originating agency, i.e., the agency that made an arrest or disposition and its county. This provides opportunities for analyses by county and law enforcement entity in combination with external data sources.

SUPPLEMENTAL DATA

The American Community Survey (ACS). For information on the state population, we used the American Community Survey by the United States Census Bureau. We accessed the data through the Integrated Public Use Microdata Series (IPUMS).[4] This dataset is released annually and is based on survey responses by individuals. This dataset allows an analysis of state and county demographic characteristics such as gender, race, and citizenship. This allows us to compare the composition of the general population with that of individuals charged with different crimes as shown in the CORI data.

Other Data Sources. When it is appropriate for the research question, we use data sources other than the population data in the ACS. For example, use of the Law Enforcement Management and Administrative Statistics (LEMAS), which is collected and released by the Bureau of Justice Statistics of the U.S. Department of Justice, allows us to conduct organization-level analyses by linking the LEMAS data with the CORI data. We may also analyze other sources of data alongside the CORI data, such as presidential electoral results; the adoption of formal criminal justice programs initiatives statewide or in a county; and legal changes such as the enactment of statewide sanctuary policies.

[1] Questions and inquiries regarding the data should be directed to Professor Michael Light (mlight@ssc.wisc.edu). This factsheet is written with the assistance of Jungmyung Kim (jungmyung.kim@wisc.edu).
[2] Research access to CORI data requires researchers to commit to a series of security-related measures. For more information, please contact the Research Center at researchrequest@doj.ca.gov.
[3] https://www.ussc.gov/guidelines/judiciary-sentencing-information
[4] https://www.ipums.org

APPENDIX 1: DESCRIPTIVE STATISTICS

Variable	# of Cases	Missing Rate[1]	Possible Values	Mean[2]	
				Before Cleaning	After Cleaning
Case outcomes					
Case Acceptance[3]	19,892,759	0.0%	0, 1	0.52	0.51
Conviction[3]	19,892,759	0.0%	0, 1	0.42	0.42
Incarceration[3]	19,892,759	0.0%	0, 1	0.28	0.28
Incarceration Months[3]	19,815,082	0.4%	0-480	3.90	4.00
Demographic characteristics					
Age	19,892,759	0.0%	[0, ∞)	33.67	33.58
Gender (Male=1)	19,891,216	<0.1%	0, 1	0.78	0.79
Race: White	19,689,103	1.0%	0, 1	0.36	0.36
Race: Black	19,689,103	1.0%	0, 1	0.18	0.18
Race: Hispanic	19,689,103	1.0%	0, 1	0.41	0.41
Race: Asian	19,689,103	1.0%	0, 1	0.03	0.03
Race: Other Race	19,689,103	1.0%	0, 1	0.03	0.03
Foreign-Born	19,817,679	0.4%	0, 1	0.20	0.18
Non-U.S. Citizen	19,356,733	2.7%	0, 1	0.16	0.16
Criminal history					
Previous felony arrests	19,892,759	0.0%	[0, 10]	4.05	4.17
Previous misd. arrests	19,892,759	0.0%	[0, 10]	3.95	4.06
Previous incarcerations	19,892,759	0.0%	[0, 10]	0.99	1.02
Case characteristics					
Number of arrest charges	19,892,759	0.0%	[1, 6]	1.74	1.75
Offense code	19,892,728	<0.1%	(0-1 indicators for 4,948 codes)		
Originating County	19,872,918	0.1%	(0-1 indicators for 58 counties)		
Number of Cases				19,892,759	19,112,520

[1] "Missing rate" was calculated by the number of cases with missing values divided by the total number of cases that had felony and misdemeanor arrests of adult defendants excluding cases that only had federal immigration law violations.

[2] The means were calculated by excluding missing values in the variables. For example, if gender of the arrestee was unknown for a case, this case was not considered for calculating the mean of the gender variable.

[3] The missing rates of these variables were calculated after treating them as law enforcement releases or case dismissal by the prosecutor. There were a few additional cases with missing incarceration length.

APPENDIX 2: EXPLANATION OF REGRESSION ANALYSIS

Regression Models. A regression model attributes the variation in outcome—"dependent variable"—to a set of explanatory factors, or "independent variables." This means that regression models explain how much the phenomena of interest (e.g., criminal case outcomes) vary depending on other factors (e.g., race, gender, and the number of prior convictions of the defendant). This simplifies the explanation and prediction of the outcome.

Controlling. Regression is a popular approach to "controlling" or "conditioning" in social science. Since there are potentially multiple factors influencing any outcome of interest, one must consider how to identify the appropriate contribution of each explanatory factor. In this way, one can estimate how much an outcome differs depending on one factor when the other factors are held constant, i.e., when the influence of the other factors is already taken into account. This allows the comparing of outcomes by the explanatory factor of interest for otherwise similarly situated people.

Statistical Significance. Analyzing data from a large population sample, the results from the sampled data could result from the randomness in the sampling process or reflect a true tendency in the population. Where a finding is considered to have statistical significance, this indicates that the empirical finding is unlikely to result from pure randomness in the sampling process. We report the existence of statistical significance to indicate substantial differences across racial groups according to the results of regression models.

Use of Regression in the Factsheets. In the factsheets, we use regression models to estimate the predicted case outcomes by race or other explanatory variables. We run the models for each outcome of interest—e.g., whether the case was charged by the prosecutor, whether the prosecution led to conviction, and whether the conviction led to incarceration. We compare each of these outcomes by race through regression.

To account for the influence of other factors on the outcome, we include a series of control variables to regression models: age, gender, and race of the arrestee, the number of prior arrests, the number of charges in the case, the year of arrest, the county of arrest, and the code of the most serious offense. This allows comparisons of outcomes among "similarly situated" individuals by race. The predicted probabilities are calculated as the average of outcomes by racial group when all the other factors are held constant at their averages.