DIRITTO PENALE
E UOMO

Riflessione

A new criminal records database for large-scale analysis of policy and behavior

by Pablo A. Ormachea[1,2], Gabe Haarsma[1,2] Sasha Davenport[2], and David M. Eagleman[1,2]

Originally published in *The Journal of Science and Law*, 2015, 1(1):1-7

Abstract. *To allow large-scale, cross-jurisdictional analyses of criminal arrests, we have developed the Center for Science and Law's* Criminal Record Database *(CRD), a collection of tens of millions of U.S. courthouse records. The CRD can enhance many types of research – for example, identification of high-frequency offenders, measurement of changes in policing strategies, and quantification of legislative efficacy – giving policy makers the best data upon which to base law enforcement decisions. The CRD provides a heightened level of detail about individual offenders, their crimes, and their interactions with the criminal justice system; additionally, it translates court records into a common framework for cross-jurisdiction comparison. In particular, the database includes anonymized identifiers to support exploration of criminal re-offense (recidivism) within the same jurisdiction. A constantly growing project, the CRD currently contains 22.5 million records from 1977 to 2014 from Harris County in Texas, New York City, Miami-Dade County in Florida, and the state of New Mexico.*

Keywords: criminology, database, big data, policy, behavior

[1] Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA.
[2] Center for Science and Law, Houston, TX, USA.

1. Introduction.

Much of adult criminal quantitative analysis relies on the FBI's Uniform Crime Reports (UCR) (Butts & Evans, 2014; Lott, 2010; Steffensmeier et al., 2011). The underlying data are voluntarily reported by law enforcement agencies across the country, and the FBI compiles the reports to publish aggregate statistics. The yearly results form one of the most comprehensive collections of violent crime and property crime in the nation. To access the data, researchers can make use of an online tool that allows research of crime statistics since 1985, and in some cases back to 1960 (FBI, n.d.). The widespread use of the UCR and other large-scale databases (e.g., the National Youth Survey and the National Longitudinal Survey of Youth) establishes the enthusiasm of the research community for longitudinal crime data.

However, for many research purposes the UCR data has some limitations:

First, because there are no unique identifiers, it is impossible to identify recidivists—that is, researchers cannot identify offenders' reentry into the criminal justice system. Second, it lacks detail about individual crimes and their outcomes (e.g., number of charges, plea bargains, dispositions, fines, jail sentences, etc). Third, records in the UCR represent cumulative figures, thereby missing interesting detail about crimes, charges, and dispositions. Moreover, only the most serious charge from a handful of possible categories is collected for the UCR's aggregate reports (FBI, n.d.). Fourth, the UCR's reliance on voluntary reports from agencies throughout the country leads to high local variance, because different jurisdictions follow inconsistent procedures and definitions of crime (Loftin & McDowall, 2010). Police complete official crime reports for the FBI at variable rates, with one researcher concluding that police made reports in only 39.3% of all violent crimes and only 49.3% of all property crimes (Loftin & McDowall, 2010). Collectively, these limitations complicate analyses.

An alternative approach to crime record analysis can be pursued by the study of individual court records, housed in hundreds of counties across the United States (Mueller-Smith, 2014). However, each jurisdiction employs local laws and sparse idiosyncratic information management systems, making it prohibitively difficult to compare detailed crime records across time and place.

To address these limitations, we have developed a database that employs tens of millions of individual criminal records from jurisdictions across the country. The Criminal Records Database (CRD) addresses the aforementioned UCR's limitations while providing a far richer collection of individual-specific data – for example, whether the defense attorney is appointed or hired, the disposition, use of deferred adjudication or shock probation, sentence length, fine amount, and more.

The advantages of this novel dataset include: (1) individual identifiers allow for recidivism analysis – albeit only for repeated bookings within the same jurisdiction, (2) the presence of all the charges allows for deeper understanding of all crime, not just a subset, (3) more and different offender-specific variables than the UCR, (4) the data represent a comprehensive and growing picture of information available to judges and prosecutors, (5) more and different disposition-specific variables, enabling assessment of small variations in punishment, (6) continual development, as we see the CRD as a data platform for the research

community, which will collaborate with us to integrate new datasets from other jurisdictions or other points in the detention process (e.g., corrections).

For all fields, the CRD bears on a fundamental dimension of human behavior: what affects how criminals make choices? By enabling an exploration of the relationships between external factors like legal policies or civic participation and the decision to commit (or not commit) a crime, we hope to enable research with results that are meaningful across and beyond the contributing disciplines. Ultimately, the CRD aims to foster scientifically based social policy by providing open-source, data-driven analysis.

2. Methods.

2.1. Data acquisition.

To acquire the underlying data, we contacted New York City (New York), Harris County (Houston), Miami-Dade County (Miami), and the state of New Mexico to obtain copies of their criminal records from their justice information management systems. As public records, the data were obtained via Freedom of Information Act requests. The Institutional Review Board at Baylor College of Medicine exempted this release of an anonymized dataset from human subject research oversight because they consist of publicly available records.

Different jurisdictions use thousands of diverse, inconsistent labels for crimes. On the low end, Harris County data consist of 3.048 million records that use 2,474 different text labels; a large proportion can be attributed to misspellings for "assault." At the other extreme, New Mexico's data consists of 3.859 million records with 5,607 different code citations. This disparity would confound even the most sophisticated automated analyses.

2.2. Data processing.

The raw data come in several different tables and in a variety of text formats with diverse – and sometimes cryptic – variable names. Our legal scholars worked side-by-side with our programmers to carefully standardize variable terminology and formatting for clarity and consistency across jurisdictions. This included conversions to a uniform date format as well as standardizing the capitalization and spellings of the field names across the different jurisdictions' datasets.

Next, we worked to eliminate obvious mistakes in the raw data entries, doing so only when we were certain that a record contained an error. In general, errors fell in two categories: first, obvious human-entry errors, as when a police officer or clerk would input 0000000 or something similar as the birth date. We used the database NULL to replace ages of 0, or impossible birth or filing dates (e.g., February 31 or births that supposedly happened after the arrest date).

The second category of errors consisted of duplicate entries. Some of these were attributable to the way the data were pulled from the information management systems (the SQL "join" command), while others were a function of the data-entry process. As an example of the latter, Harris County enters the end of probation (successful or unsuccessful) as a new, duplicate record. Instead of treating it as a separate event, we transfer the information to disposition of the original record. Similarly, offenders who have been assigned multiple identifiers from different

states often had multiple record entries for precisely the same offense. Another source of duplicates arose occasionally from revisions to prior cases (e.g., updating the postal code or correcting a simple clerical error). In all cases, we erred on the side of caution and only removed duplicates when certain of the underlying reason and able to verify it with the appropriate county or state clerks.

2.3. Designing the categorization system.

To address the issue of diverse, idiosyncratic charges across jurisdictions, three lawyers created and implemented a categorization schema, working closely with the team's programmers to understand and classify the thousands of different crime labels across multiple counties in different states. We sought to design a novel categorization system broad enough to overcome the subtle variation in elements of an offense, yet detailed enough to allow for a granular understanding of how different crime types have changed over time.

To allow for different research purposes, we designed two levels of categorization: *Broad* and *Detailed*. The *Broad* categorization contains 32 classifications ranging from theft, to murder, to crimes by public servants. The *Detailed* categorization divides the same data into 152 more fine-grained classifications, ranging from a second time DWI, to social services fraud, to possession of a small amount of marijuana. See Supplementary Material for details of both categorization schema.

All crimes were then assigned within the two categorization schemes. For example, a crime of using a bad check in New York City would be labeled "Fraud/Forgery/Impersonation" (*Broad*) and "Theft – Check" (*Detailed*), depending on the appropriate penalty. We followed current delineations in the legal code wherever possible, but the discrete nature of categorization necessitated the prioritization of certain crime types over others. In those instances, we were guided by the offense's typical severity as well as the underlying motivation of the crime. For example, if a public servant were arrested for homicide, we would place that crime record under homicide instead of crimes by public servants. Our hierarchical schema for categorization is detailed in (**Fig. 1**).

Next, to process the raw data into a common currency, we identified and surmounted three types of obstacles. First, the datasets varied in terms of data entry. For example, Harris County uses text entry fields into which police officers often type different spellings for the same crime. Assault, for instance, is commonly entered as "asslt", "assault", "aslt", and several other variants. New York and Miami use thousands of references to the criminal code which our lawyers manually identified, evaluated, and assigned to the appropriate category. Deciphering these abbreviations and identifying the code citation required extensive collaboration between our programmers and lawyers.

Figure 1. A hierarchical schema for classifying crimes into categories

Second, labels change over time and place. As an example from Harris County, the term "rape" was replaced by "sexual assault" in 1985 (**Fig. 2**). Without this knowledge, and a conversion of the terms into a common currency, any automated analysis is unlikely to succeed. Sexually violent crimes also require different legal elements in different jurisdictions. In New York, the crime of sexual abuse simply requires "sexual contact" without consent. This is a fundamentally different evidentiary burden than, for example, Miami's sexual battery which requires the involvement of a sexual organ. Our analysis of legal codes also showed that one location can criminalize different behavior without a precise analogue in another. For example, New York and Miami-Dade County have specifically criminalized genital mutilation of a female child. Harris County has never prosecuted this specific crime, yet would likely classify it under sexual assault of a child. Again, without this knowledge and the subsequent comparisons it allows, a cross-regional analysis would be impossible.

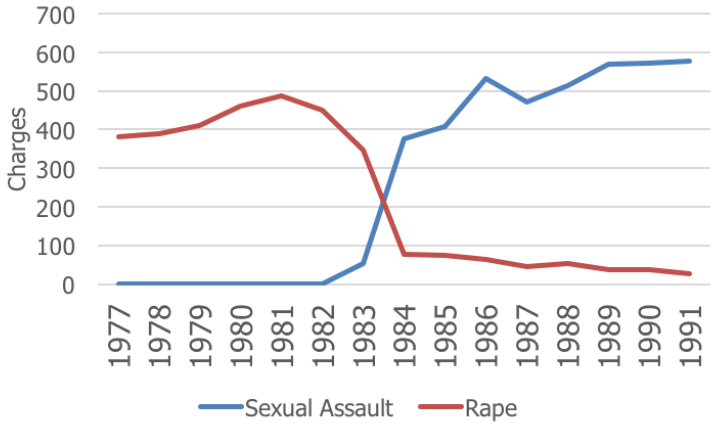

Figure 2. In Harris County, TX, the term "rape" was replaced by "sexual assault" in the early 1980s. Without a classification system that bins both labels into a common category, an algorithmic analysis would erroneously conclude that rape arrests in Harris County had come to an end.

Third, underlying crime definitions have changed. For controlled substances, different jurisdictions spent much of the 1980s identifying drugs by their specific name, such as "crack" or "methamphetamine." Today, however, nearly all controlled substances are classified according to charts grouping drugs by associated penalty. We have thus lined up those older crimes with their modern day counterparts to allow for analyses across time.

To illustrate the process of crime categorization, here is an overview of the process for sexual crimes: We first evaluated 3.1 million individual criminal records from Harris County since 1977 and identified the relevant subset of records involving sexual offenses (n=132,099). Next, we sorted those records into three categories: sex crimes committed against minors (n=32,819), sexually violent crimes (n=10,177), and nonviolent sexual crimes (n=89,163). Finally, the three broad categories were divided into detailed subcategories ranging from sexual abuse of a child, to indecent exposure, incest, or burglary with attempt to commit a sexual assault. The sexually violent crimes, for example, include 10 subcategories and maintain a distinction between one-time (sexual assault) and repeated sexual violence (sexual abuse) – distinctions that would be lost in the UCR.

We designed the *Broad* categorization to enable cross-comparison across time and place. Moreover, those researchers interested in particular types of crime (e.g., white collar) can use the schema to identify the subset of charges relevant to their research (e.g., "Fraud/ Forgery/Impersonation" and "Theft") and thereby avoid combing through the entirety of CRD. It is critical to note, however, that the *Detailed* categorization only provides researchers with additional detail if the raw record itself allowed us to place it in a discrete category. It would therefore be improper to compare the *Detailed* subcategories from one jurisdiction to those of another. In Harris County, TX, for example, the record labels within the *Broad* category of "theft" do not allow us to distinguish generic theft from theft of transportation. In contrast, New York City's statute citations do make this distinction in the records, allowing subcategorization of the offense. A comparison of this subcategory in NYC against Harris County would wrongly conclude that Houstonians are never arrested for theft of transportation. In summary, the *Detailed* subcategorization is intended as a tool for diving deeper within individual jurisdictions.

We recognize that no single classification scheme will suit the needs of all researchers; to that end, researchers interested in downloading the data will receive the raw labels as well as

our default categorization schema[3].1 We also provide the raw labels in the hopes that the research community will help evaluate our categorization schema. Wherever possible, we have developed libraries and graphical interfaces to simplify the end-user's role in making these choices. For outside researchers relying on the CRD, our terms and conditions require them to specify the categorization schema they are using, whether the default or a clear description of their modifications.

2.4. Designing the categorization system: Dispositions.

We next converted the dispositions into a common currency, as the different jurisdictions have as many as 250 slightly different labels to describe the conclusion of the criminal process. Such labels can range from the simple (e.g. "dismissal" or "guilty plea") to the more unusual (e.g. "diversion" or "shock probation"). Our team manually evaluated each disposition and sorted them into one of six categories. The first four categories are "dismissed," "acquittal," "guilty," and guilty by plea." The fifth category, "conditional dismissal," covers those cases in which the accused is functionally guilty but able to automatically remove the crime from his record after fulfilling certain conditions. These conditions typically include a fine and community service as well as a probationary period. In Texas, this is typically known as deferred adjudication. In New York, this is known as adjournment in contemplation of dismissal. The final category, "No Action," appears when a charge was filed but the district or county attorney declined to prosecute. Within our datasets, this disposition is most common in Miami-Dade County and least common in Harris County, Texas.

2.5. Added fields.

Although the CRD is comprised of public records, we have taken steps to minimize potential invasions of privacy by de-identifying the individual records. Our anonymization process uses the official identifying number as well as an MD5 one-way hash to generate a 128-bit variable. This approach generates a final object for which it is nearly impossible to determine the original object. Collisions are mathematically rare; to ensure there were none, we verified that the number of unique identifiers remains the same before and after the anonymization. Finally, we leave the birth month and year but remove the day to further protect individual privacy. In our curated, public-facing dataset, we only make available this anonymized data, stripped of individual identifiers. Internally, we maintain a file that links the de-identified data with the subject names and identifiers[4].

We have also enhanced the original data by adding new fields. For example, the raw data in some jurisdictions only identifies an individual's ethnicity as black or white. Because it is likely that many Hispanics are being incorrectly identified as one of those two races, we follow the methodology of the U.S. Census Bureau to estimate Hispanic ethnicity from surname (Perkins 1993). Our algorithm imputes values even where the race field is already populated, and places

[3] The internal categorization process relied on a human-entry process. The different jurisdictions in the database (currently New York, NY; Harris County, TX; Miami-Dade County, FL; and the state of New Mexico) contain a total of 13,398 different labels or code cites. For instance, even a 0.1% error rate – which is on the lowest end of available research on human-entry error rate (see Panko n.d.) – would still result in 14 classification errors. We have been as careful and as thorough as possible, but it is still likely that there are minor slips. Therefore, CRD deliberately includes the raw labels and look-up tables not just for people to adjust labels for their own research purposes but also to engender discussion that improves the CRD for the entire community).

[4] Researchers seeking to make use of the identifiable data should contact the Initiative to explore opportunities for collaboration. Such projects would require IRB review and approval.

the results in a new field in the database. This means the CRD includes the original race variable as well as an additional, inferred race variable. Although this method of estimating Hispanic heritage cannot be perfect, it is considered to have minimal false positives (Perkins, 1993). This methodology is most effective in estimating the number of Mexican Americans or Puerto Ricans, but it is known to undercount people of Cuban or "Other Hispanic" descent (Perkins, 1993). This undercounting is most noticeable in regions with sizable populations of Cuban or South American descent, such as Miami-Dade County. This method also undercounts individuals who would identify as half-Hispanic, because the last name would change if one of the parents had taken the other's last name.

Similarly, we used U.S. Census-derived tables to infer gender. Here, however, we only imputed a gender where the record was missing. As we do elsewhere, we erred on the side of caution as follows: we took from the U.S. Census the top 1,200 names for men and the top 4,275 names for women. Male and female names have different distributions, with 60 names covering 50% of the male population but with 139 names covering 50% of the female population. Given the variation, we used different cut-offs: for men, the male-female census ratio needed to exceed 10:1 to be used to infer male gender. For women, the female-male ratio needed to exceed 5:1 to be used to infer female gender.

3. Discussion.

We have developed the largest, open-source, comprehensive, and de-identified database of 22.5 million criminal records, spanning 35 years from 1977 to 2012. This resource opens the door to an array of research questions—for example an analysis of the outcomes of those who plea versus those who do not, analyses of the number of times a person pleas, the impact of different combinations of prison time and fines on re-offense rates, and so on.

The inclusion of an anonymous identifier enables a deeper understanding of criminal re-offenses. Improvements in the ability to identify individuals responsible for repeated bookings within a jurisdiction could inform evidence-based policies aimed at the prevention and control of crime. This will benefit society at-large, allowing policymakers to base law enforcement decisions like sentencing or allocations of funds on quantitative assessments of criminal propensity.

As one of its strengths, the CRD enables a cross-jurisdictional comparison of the criminal arrest process. For example, Miami-Dade County has had 1,906,298 arrests with a disposition of "No Action," which means the district attorney opted against prosecuting the crime. Over the same time period, Harris County had only 27,029 arrests with the same disposition. We attribute this stark difference to Texas' requirement that police officers call the district attorney to preauthorize arrests. Understanding differences like these can translate into enormous cost-savings for jurisdictions.

Along with its strengths, the CRD also has limitations. We turn to these now, with potential solutions where possible.

1. The jurisdictions currently represented in the CRD do not identify offenders of Hispanic descent. To obtain a better understanding of the demographics, we have estimated the Hispanic population by last name.

2. The database contains no juvenile records, as those are not included in basic Freedom of Information Act requests. We note that juvenile is defined differently in each locale, so 17 year

olds are included in Harris County records whereas only 18 year olds appear in New York City and Miami-Dade County records.

3. The database does not include sealed or expunged records, as those are typically removed from the underlying county databases. It is likely that this disproportionately affects certain crime types (e.g., traffic offenses).

4. The CRD does not have victim data, precluding analysis of, for example, whether ethnicity or age of victim affects sentencing.

5. All the records in the database were originally entered by humans. Aside from typographical errors (which were relatively straightforward to fix), a larger problem is missing data. For example, some fields have become more populated with time. Birth date was not as commonly entered in some of the earlier records from the 1970s and 1980s, but becomes more rigorously entered with time (**Fig. 3**).

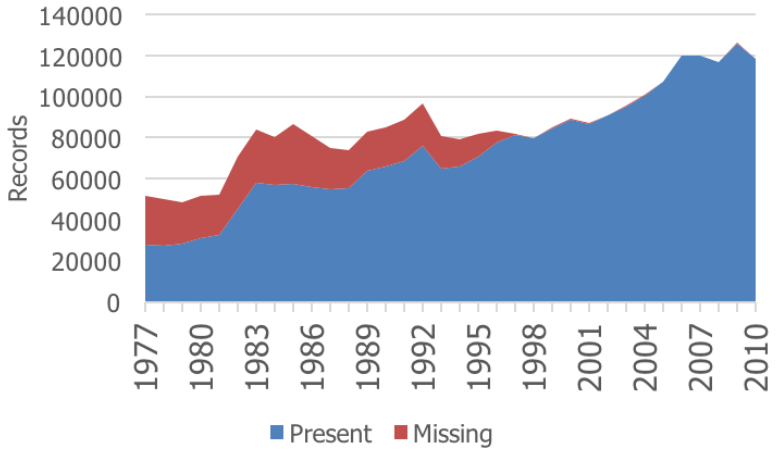

Figure 3. Birth dates in records in Harris County. The stacked histogram demonstrates that many variables (date of birth, in this case) become more completely populated in the records over time**.**

6. The CRD does not contain corrections records, as most states do not consider those public. Therefore, while we know each offender's sentence at the end of trial or plea bargaining, we cannot know how long an offender actually served. This is potentially solvable by marrying CRD data with independently obtained corrections records, a strategy we are currently pursuing.

7. While our *Broad* categorization allows for comparisons across jurisdictions, our *Detailed* categorization does not: the subcategories become populated only if the jurisdictions' labels or code citations provided enough detail.

8. Some jurisdictions have more limited data than the rest. For example, New York City's records only list the most serious offense per arrest and do not yet include an identifier. We are currently working to obtain the missing data for NYC.

9. There is some state-by-state variation in terms of privacy. Currently, the CRD does not contain data from the Northwest and Midwest, as those states have stricter privacy laws for criminal records.

10. The recidivism analysis allowed by the CRD only applies for repeated bookings within the same jurisdiction. This approach will systematically undercount the true recidivism rate due to relocation.

11. The CRD only contains arrest data and not incident based data, thus providing a picture of crime at the courthouse level. This means that previous stages in the law enforcement process (e.g., 911 calls, house calls, etc.) could skew the arrests that make it into courthouse databases. In contrast, the UCR includes all reports to law enforcement, providing a different angle on criminal activity.

Despite these limitations, the CRD can serve as an open-source resource for the research community, providing a large and detailed database for cross-jurisdictional comparisons. Downloads can be accessed at http://www.neulaw.org/data.

ACKNOWLEDGEMENTS

This work was funded by the National Science Foundation grant #11439453 and a charitable grant from Nicholas and Susan Pritzker. For extensive help throughout the processes of categorization, we thank Sarah Isgur Flores, Leah Jorewicz, Jayme Reisler, Joshua Preston, Frances Harvey, and Sameer Birring.

WORKS CITED

A. Butts, D.N. Evans, *The Second American Crime Drop: Trends in Juvenile and Youth Violence*. In W.T. Church II, D. Springer, A.R. Roberts (Eds.), *Juvenile Justice Sourcebook*, Oxford University Press, 2014, pp. 61 ss.

Federal Bureau of Investigation (FBI), United States Department of Justice. (n.d.). *Uniform Crime Reporting Statistics*. Retrieved from http://www.ucrdatatool.gov.

C. Loftin, D. McDowall, *The use of official records to measure crime and delinquency*, in *Journal of Quantitative Criminology*, 26(4), 2010, pp. 527 ss.

J.R Lott, *More Guns, Less Crime: Understanding Crime and Gun Control Law*s, University of Chicago Press, 2010.

M. Mueller-Smith, *The Criminal and Labor Market Impacts of Incarceration* (Working Paper), 2014, Retrieved from Columbia University website: http://www.columbia.edu/~mgm2146/incar.pdf.

R. Panko, R. *Ray Panko's Human Error Website*. Retrieved from http://panko.shidler.hawaii.edu/HumanErr/Index.htm.

R.C. Perkins, *Census Bureau, United States Department of Labor*. *Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results*, 1993, retrieved from http://www.census.gov/population/www/documentation/twps0004.html.

D. Steffensmeir, B. Feldmeyer, T. Harris, T. Ulmer, *Reassessing trends in black violent crime, 1980-2008: Sorting out the "Hispanic effect" in uniform crime reports arrests, national crime victimization survey offender estimates, and U.S. prisoner* counts, in Criminology, 49(1), 2011, pp. 197 ss.