FINAL REPORT

OCTOBER 2021

Improving Data Infrastructure to Reduce Firearms Violence

Chapter 5. Expanding Capacity and Capabilities to Monitor and Research Guns in the United States

Michael Mueller-Smith, PhD | University of Michigan

Editors:

John K. Roman, PhD
NORC at the University of Chicago

Philip Cook, PhD
Duke University

Chapter 5. Expanding Capacity and Capabilities to Monitor and Research Guns in the United States

Michael Mueller-Smith, PhD | University of Michigan

Introduction

As information technology systems and survey operations have modernized over the last half-century, a range of novel data collection opportunities have emerged, advancing our ability to track and measure various forms of socioeconomic activity and outcomes critical to U.S. public policy. Numerous examples of cutting-edge data infrastructure have been founded and developed from administrative records not originally produced for research purposes but instead a byproduct of regular day-to-day operations of individuals, governments, nonprofits, and businesses. Examples can be found in a variety of policy domains: labor markets (Longitudinal Employer Household Dynamics program); education (National Student Clearinghouse); criminal justice (CJARS); and health care (HCUP). Such systems have successfully navigated a range of serious legal (e.g., HIPPA, FERPA) and privacy hurdles to provide major advances in data infrastructure, a prerequisite for building evidence to inform policy.

At the same time, survey data collection efforts are modernizing in response to changing conditions. Internet-based outreach efforts (including operations like Amazon MTurk) have lowered the cost of conducting surveys and expanded the possibility for new types of experimentation with subjects, with the caveat of having a nonrepresentative sample. Additionally, traditional representative surveys are gaining new life as expanded research possibilities have emerged for individual-level linkages with other survey and nonsurvey datasets, increasing the range of questions and research designs that can be explored with the data.

Leading efforts at data architecture integrate both survey and administrative data sources to flexibly approach data collection, capitalize on respective strengths, and minimize potential weaknesses. In the context of studying gun ownership, gun use, and its effects on the population, this is even more critical given the non-trivial legal and political barriers to progress. What is needed is a multifaceted strategy of interoperable, diversified, complementary collection efforts that together generate a sum greater than its parts.

Highlighting a model data system: The CJARS, founded in 2016, is an ongoing data collection effort and cutting-edge dissemination platform, designed to transform research and statistical reporting on the U.S. criminal justice system. It is the first nationally integrated research repository that follows individuals from arrest to charge to disposition to sanction.

Data come from all types of criminal justice agencies and from across the United States. At the University of Michigan, data are harmonized into a common schema that allows analysis across disparate jurisdictions. After secure transfer to the U.S. Census Bureau, CJARS data are anonymized and linked at the individual-level to confidential social, economic, and demographic survey and administrative records to produce novel empirical analysis of criminal justice caseloads.

The project's ultimate goal is to enable research and statistics that legislators and administrators can use to develop evidence-based criminal justice policy. Data from 23 states are currently held.in the U.S.

Targeted efforts should tackle specific measurement goals without trying to solving all policy-relevant questions, providing a viable path forward that limits the opportunity for individual barriers to halt progress. Examples include:

1. Purchase or transaction data that may cover federally authorized gun dealers, including credit card transaction data
2. Health records (usage or claims) that indicate gunshot victimization
3. Mortality records that identify potential gun involvement
4. Weapons offenses in criminal justice records
5. Embedding gun-related questions (for instance, current and prior ownership) into ongoing nationally representative survey efforts: National Longitudinal Survey of Youth, General Social Survey, or the Children's Health Survey
6. Engaging modern survey platforms for social science research to assess willingness to pay

Central to the success of this approach is intentional planning to support linkage, such that any individual success builds broader momentum. Key linkage factors that should be considered include:

1. Individual personal identifiers or other personally identifying information variables
2. Unique gun identifiers
3. Local geographic information where applicable

Integrating multiple data sources produces timely, policy-relevant evidence: In response to the COVID-19 pandemic, Congress created the Paycheck Protection Program (PPP) to support small businesses. Original provisions from the Small Business Administration (SBA), however, made businesses ineligible for the PPP if an owner had a variety of recent contact with the justice system. Using secure data infrastructure in the Federal Statistical Research Data Center network, Finlay, Mueller-Smith, and Street (2020) investigated these restrictions using individual tax return data linked at the person-level with CJARS-covered criminal records. They found that as many as 3.2 percent of sole proprietorships may have been ineligible for PPP assistance due to current or prior criminal justice involvement. Black and Hispanic men with sole proprietorship income were significantly more likely to be PPP-ineligible than white men. Between 6.9 and 15.4 percent of former convicts rely on self-employment income, which is particularly pronounced for women.

In part due to this evidence, SBA later relaxed these provisions, expanding access to PPP support for over 1 million entrepreneurs with criminal histories.

Scaling Infrastructure With Machine Learning

While administrative data have brought significant promise to several research and policy domains, they introduce serious challenges because the underlying information was produced for operational, not research, purposes. A common problem is the existence of free entry text fields or case notes that contain a wealth of information but typically require human review to extract relevant information for analysis purposes. When reviewing thousands if not millions of records, data harmonization becomes infeasible.

Some existing data collection efforts have invested significant time and resources in hand-coding raw data fields to formalize research schema. For example, the NEISS trains human coders to differentiate between “accidental,” “assault,” and “self-inflicted” injuries from free text fields, which is thought to substantially address overuse of the “accidental” classification. Other data resources are ripe for further investigation. For instance, national crime reporting systems like the Uniform Crime Reports or the NIBRS fail to differentiate between crimes involving gunshots versus gun threats, which some localities already differentiate and could be expanded through broader use of case notes for data processing and preparation. At scale, hand-coding records through human review for classification purposes becomes cost-prohibitive in many cases, limiting the available information from extant records for analysis purposes.

Recent years have seen an explosion of activity in machine learning and data science, which potentially provide a cost-effective path forward. Existing data series provide a rich source of *training data* from which machine learning models can be constructed. This would reduce the financial costs of operating those data collection efforts going forward, as the estimated models can be used to algorithmically classify unambiguous records, focusing more resource intensive human review on ambiguous entries.

Machine learning also generates broader returns as complementary collection efforts may seek to analyze similar and overlapping field content at the state and local level. The production and broad distribution of such trained machine learning algorithms can empower local actors to leverage subnational data, which often entail lower access barriers compared to national datasets, and lower research costs overall. It also provides a common framework to develop estimates based on the same underlying definitional concepts, avoiding the problem of apples to oranges comparisons when different organizations use different classification criteria.

Machine learning in practice to advance data infrastructure: Currently, CJARS is built on over 2 billion lines of raw data, covering approximately 178 million unique criminal justice events, occurring in 23 states. Most agencies provide free entry text fields to describe the type of offense involved in a given criminal episode, resulting in over 4 million unique offense descriptions in the data. Choi, Kilmer, Mueller-Smith, and Taheri (2021) leverage a unique source of 386,906 classified offense descriptions produced by Measures for Justice (MFJ) to train a machine learning algorithm, known as the Text-based Offense Classification (TOC) tool.

TOC is now used by both CJARS and MFJ to support quality data processing at reduced costs. It also supports the research and analysis of a range of external organizations and will be launched in late 2021 as a public tool for all to use.

Disciplining Data Construction to Avoid Bias

Administrative records hold significant promise for research and statistical purposes, but one must confront the fact that they are a byproduct of operational uses and not originally designed for analysis. This creates a number of unique challenges, including data and validation. In addition, researchers must have a clearly defined target observation unit for the intended data product and strategies to translate one or more source files into that structure.

To take a simple example: suppose a researcher were working with payment data and saw a series of regular monthly transactions with a licensed gun dealer. Should these be interpreted as distinct purchases indicating multiple gun transactions, or do the combined payments represent a single purchase being paid off in a monthly payment plan? What if there is a very large single purchase? Should this count as a single gun transaction or multiple? The answer depends on how one defines the intended unit of observation and the viability of defining a mapping from the source data to that target.

We can think of a second example in the health context: suppose a patient has multiple health events at a hospital over the course of several weeks that have been tagged as being associated with gunshot wounds. Ultimately, the patient dies, which creates an additional death record that also notes a gunshot wound. Does this constellation of events represent a single shooting event leading to immediate care, follow-up care for complications, and ultimately death? Or, are the events distinct gunshot events for an individual in a crisis period?

While these examples might seem contrived, decisions on how to handle such situations will have fundamental implications for measurement. In the payment data example, gun prevalence rates could be dramatically over- or underestimated. Similarly, in the health example, the number of victimization events could be three times too high or one-third the true rate.

Such problems become even more complicated when combining multiple dataset sources, whether across non-mutually exclusive jurisdictional boundaries (e.g., a gunshot victim seeks treatment at both an urgent care facility and a hospital emergency room) or from the same provider over time (e.g., a health care event that was previously nonfatal is later reclassified as fatal).

Two features are critical to navigating these problems:

- Defining or developing unique identifiers that combine related events or observations from the same individual, household, or gun
- Implementing a strategy to disambiguate or deduplicate records down to the intended unit of observation for the target dataset (a process that can be disciplined by validation)

Benchmarking CJARS against federal statistical series: The United States lacks uniform rules across state and local jurisdictions on the privacy afforded to justice-involved. Likewise, there is substantial variation in the development of data access mechanisms for researchers. Lacking authority to compel data provision, CJARS relies on multiple strategies for opportunistic data acquisition, including data use agreements, public records requests, web scraping, bulk data downloads, and data donations. Data arrives in provider-specified formats and structures, which then have to be reconciled by staff at the University of Michigan. Due to the variation in data collection methods and the numerous creative solutions required to coherently process the data, there is a fundamental need to benchmark CJARS against other available data series to identify both the strengths and weaknesses of CJARS.

Papp and Mueller-Smith (2021) report the ability of CJARS to reproduce a range of statistical series published by the BJS, including the State Court Processing Statistics (SCPS), National Prisoners Statistics (NPS) Program, National Corrections Reporting Program, Annual Probation Survey, and Annual Parole Survey. Such comparisons have enabled CJARS to identify shortfalls in its data processing, and improve the quality and accuracy of the data product.

Benchmarking Strategies to Validate Data Quality

A multifaceted approach to data collection provides flexibility and agility in response to changing legal and regulatory environments across jurisdictions and over time. For instance, combining information on transaction data, permit records, and self-identified ownership in survey responses could provide one of the most accurate measures of gun prevalence in the United States, collectively addressing the individual measurement and attrition biases of any individual source in isolation.

But, it also creates a number of serious challenges, including 1) numerous distinct native data layouts, especially if collecting information from state and local sources with locally defined data structures and formats; 2) inconsistent variable definitions, value codes, and free entry fields for categorical variables; 3) inadequate unique identifiers; and 4) potential duplicative coverage when receiving data from multiple providers with overlapping jurisdiction or when receiving multiple rounds of data over time from the same source.

Machine learning approaches previously described can help manage these types of data integration efforts operate at scale. Still, some tasks require tailored human engagement to understand the nature and content of a given data file.

A resulting data product is the consequence of a multitude of discretionary choices. Without an organizing framework to guide these decisions, the end result likely does not deliver on its promise.

In the context of a related data infrastructure effort, CJARS linkable person-level criminal justice records have been validated through replicating extant aggregate statistical series, including the SCPS, the NPS, and the Annual Probation and Parole surveys. Benchmarking aggregate information produced from CJARS microdata against accepted aggregate reporting programs achieves two goals. First, it provides a framework to guide data processing decisions. Second, it provides a benchmark against which to gauge data quality.

In the context of gun ownership, gun use, and its effect on the population, a number of plausible statistical series could be leveraged for benchmarking purposes to validate a new micro dataset, including:

1. General Social Survey (GSS)
    - Times series variation in national prevalence of gun ownership among households since the 1970s

2. RAND State-Level Estimates of Household Firearm Ownership
   - State-level ownership estimates (1980-2016) built from integrating survey data sources (Behavioral Risk Factor Surveillance System [BRFSS], Gallup, GSS, and Pew Research Center) with administrative data from firearm-involved suicide rates, hunting licenses per capita, magazine subscriptions to *Guns & Ammo*, and the number of background checks from the National Instant Criminal Background Check System
3. NVSS
   - Gunshot fatalities over time and across geography
4. FBI's SHR, NVSS, NVDRS, Fatal Force database
   - Law enforcement use of force across geography and over time
5. NEISS
   - Nonfatal gunshot injuries

Diversified Access Mechanisms to Balance Research, Privacy, and Security

A successful data platform should embrace multiple access mechanisms to serve multiple stakeholders.

At one end, consider the secure and confidential integrated microdata environment of the Federal Statistical Research Data Center network. This platform has been created to support research and statistical analysis of integrated data across multiple content domains and information owners. Examples could include:

1. Earnings trajectories before and after victimization events
2. Family and peer social spillovers in gun ownership rates
3. Victim/offender overlap between victimization and offender criminal justice records

While such a data environment provides the most significant promise for pushing the frontier of knowledge in this area, access and approval involve significant barriers, including potential financial costs, federal background checks, and physical limitations on where research can be performed.

Two complementary approaches have helped navigate the balance between privacy concerns and information availability. The first is a public data portal that curates aggregate statistics generated from linked microdata. The goal here is to remove any individual information, and with enhancements from differential privacy, protect the confidentiality of individuals covered in the data. Putting such information in the public domain increases transparency and data access, especially for efforts like evidence building to support data-driven policy, which often does not require individual-level information.

Second, for questions that aggregate statistics may not thoroughly answer, a synthetic data product that replicates the underlying variation in the confidential microdata but is artificially generated provides another avenue to lower access barriers without compromising privacy and confidentiality.

Balancing availability and security in practice: CJARS represents a significant (albeit growing) advancement in studying the U.S. criminal justice system. With support from the National Science Foundation, it is developing and deploying two modes of data access. First, secure access for qualified researchers on approved projects is supported through the Federal Statistical Research Data Center system. This access mechanism provides researchers with the ability to study confidential (anonymized) microdata that can be integrated with a range of other survey and administrative data held by the federal government.

In addition, CJARS is developing a synthetic data product composed of artificial records that preserve the underlying statistical information contained in the CJARS microdata. This will be publicly available via the University of Michigan's ICPSR. While the latter mode of access does not provide as many opportunities, its low-barrier approach will encourage broader adoption of CJARS in research and analysis.

Concluding Thoughts

Evidence-based policymaking is a grounding principle of good governance. In the domain of guns and gun violence, there remains a striking gap between research capacity and societal impact. Lives and communities are transformed when shootings occur, yet the absence of critical data infrastructure required to study these topics renders political discourse unmoored and unproductive, lacking a range of empirical facts to guide debate.

The United States needs a new approach to address these shortcomings. Complementary survey and administrative data collection efforts should be engaged, with diverse strategies to ensure against single points of failure. Lessons from recent advances in data science and machine learning should be embraced to reduce production costs while enhancing data quality. Intentional capacity for interoperability will build and sustain momentum as these projects mature. Together, these efforts will promote improved evidence building capacity and support the adoption of policies that enhance the safety, productivity, and well-being of communities across the United States.

References

Finlay, K., Mueller-Smith, M., Street, B. 2020. "Criminal Disqualifications in the Paycheck Protection Program." U.S. Census Bureau Work Paper Number ADEP-WP-2020-04.

Choi, J., Kilmer, D., Mueller-Smith, M., Taheri, S. 2021. "Hierarchical Approaches to Text-based Offense Classification." Working Paper.

Papp, J., Mueller-Smith, M. 2021. "Benchmarking the Criminal Justice Administrative Records System's Data Infrastructure." Working Paper. https://cjars.isr.umich.edu/benchmarking-report-download/.