Title:
Unlocking Data to Improve Public Policy
Subtitle:
When properly secured, anonymized, and optimized for research, administrative data can be put to work to help government programs better serve those in need.

Authors:
Justine S. Hastings[1,2,3,4], Mark Howison[1,2], Ted Lawless[1], John Ucles[1], Preston White[1]

Author affiliations:
[1] Research Improving People's Lives, Providence, RI, USA
[2] Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA
[3] Department of Economics, Brown University, Providence, RI, USA
[4] National Bureau of Economic Research, Cambridge, MA, USA

Words:
3,103

Key Insights

- Fact-based policymaking – the practice of using data and research to guide policy decisions – is a promising solution to improving the effectiveness and efficiency of government programs.
- Administrative data can provide new facts to guide policymakers. However, understanding the quality of administrative records, and integrating, transforming, and optimizing them for policy insights presents many challenges.
- To overcome these challenges, we developed an integrated database of administrative records from multiple Rhode Island state agencies with over 800 tables and 2.7 billion records related to over 4 million anonymous individuals. These data support econometric and machine-learning research into policies with promise to deliver higher impact per dollar and better serve Rhode Island families.
- As a specific example, we describe how anonymized data from the integrated database are used to discover a new insight into a policy challenge related to low birth weight newborns.

Introduction

There is a growing consensus among policymakers that bringing high-quality evidence to bear on public policy decisions is essential to supporting the effective and efficient government that their constituencies want and need. At the U.S. federal level, this view is reflected in a recent congressional report by the Commission on Evidence-Based Policymaking, which recommends creating a data infrastructure that enables "a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy" [4].

This article describes a new approach to data infrastructure for fact-based policy, developed through a partnership between our interdisciplinary organization Research Improving People's Lives[1] and the State of Rhode Island [13]. Together, we constructed *RI 360*, an anonymized database that integrates administrative records from siloed databases across nearly every Rhode Island state agency. The comprehensive scope of *RI 360* has enabled new insights across a wide range of policy areas, and supports ongoing research into improving policies to alleviate poverty and increase economic opportunity for all Rhode Island residents (see Sidebar #1). Our approach can guide other policymakers and researchers seeking to similarly transform and integrate administrative data to guide and improve policy.

Sidebar #1: Policy areas in which *RI 360* has contributed insights

- Lowering non-urgent emergency health care costs
- Curbing the opioid epidemic
- Improving worker training programs
- Creating tools to connect dislocated workers to benefits
- Helping families become more food secure
- Optimizing energy policy for low-income families
- Helping children reach proficiency on reading and math tests
- Closing the college achievement gap

The role of administrative data in policymaking

Administrative data can be collected from the computer systems used by government agencies to run their programs. When transformed into databases that are more suitable for insights, these anonymized records provide new sources of facts for policymakers to benchmark goals and measure the successes and shortcomings of existing and future programs. Often classified as "big data" due to their volume, variety, and availability [10], administrative records are also an increasingly valuable source for empirical social science research [5]. Research with administrative records can contribute new data-driven insights to inform important policy decisions (see Sidebar #2), and add objectivity and scientific rigor to measuring program impact and designing effective program changes. Moreover, scientists can inform how data from administrative systems, which are primarily designed around operational needs and often not suitable for analysis, can be transformed effectively to support research and insights.

Sidebar #2: Recent Data-Driven Insights from Administrative Records

- Records from the New York City criminal justice system show how judges often mispredict risk when making bail decisions [15]. Judges identify and release many defendants who have a low flight risk, but also release nearly half of the defendants with the highest flight risk. In simulations, replacing judges' decisions with a machine-learning

[1] https://ripl.org

prediction can reduce either crime rates (at a fixed jailing rate) or jailing rates (at a fixed crime rate), and in both cases can reduce racial disparities in outcomes.
- Transaction data from a private grocery retailer and data from the Supplemental Nutrition Assistance Program in Rhode Island show that households treat their nutrition benefits as if they were earmarked for food expenses, even when they could be substituted for cash [14]. This finding contradicts traditional economics theory which predicts that nutrition benefits should be fungible (e.g. substitutable for cash), and instead supports an alternative economics hypothesis called mental accounting. Results suggest that Supplemental Nutrition Assistance Program impact on spending and nutrition can be influenced by policies governing when and how benefits are distributed.
- Federal income tax records show there are growing inequalities in life expectancy in the U.S. across socioeconomic factors [2]. The breadth and scale of these administrative data (with over 1.4 billion person-year observations) reveal that geographic factors like government expenditure and fraction of immigrants and college graduates are positively correlated with life expectancy at the bottom of the income distribution.
- Randomized field experiments in Chicago combined school and unemployment insurance records with arrest records to evaluate the impact of a summer job support program for youth [6]. By using this integrated administrative data, the study found that the program caused declines in violent-crime arrests even though there were no significant effects on school or employment outcomes, which are the more typically studied effects of youth job programs.

Although the idea of guiding policy with data dates back to the 70s and 80s, early studies only considered isolated data sources and come from a time when data were scarce. It was not until recently that advances in data collection, storage, and scale provided the opportunity to integrate data across nearly every facet of government. Early case studies and survey studies highlight how the process of data modeling can facilitate negotiation and consensus-building among policymakers [8], but also how the unmet promises of new information technologies prompted frustration among government leaders at that time [9].

An important lesson is to engage policymakers and leaders to fully understand their needs, which is why we formed extensive partnerships with state government leaders while building *RI 360*. Integrated administrative data can support not only academic research, but also the analytics needs of government itself. Like researchers, government analysts need access to data that have been transformed to provide insights and integrated across programs that serve what are often overlapping populations. For these reasons, *RI 360* was selected as the primary data source for the Rhode Island Executive Office of Health and Human Service's Data Ecosystem project, to empower their data analysts and partners with data optimized for insights.

An example policy for low birth weight newborns

Throughout this article, we will describe our process for building *RI 360* in the context of a specific policy: determining the optimal weight threshold for providing additional medical care and resources to low birth weight newborns and their mothers [3]. Children born with low birth

weight tend to have more health difficulties and worse outcomes later in life compared to their peers. They also tend to be at higher risk, coming from disadvantaged backgrounds where mothers are more likely to be teen mothers or have reported alcohol or drug abuse. Programs to support these infants and mothers may increase equity of opportunity and reduce state and federal expenditures for support programs and anti-poverty programs later on in life. Currently the threshold for additional resources is set at 1,500 grams [1]. We use this threshold to measure the causal impact of these additional resources to determine if increasing this threshold could be a low-cost, high-return policy change that could improve lives, increase equity of opportunity, and save state and federal funds in the long run.

Using integrated data from *RI 360*, we can examine a wide range of outcomes, including educational test scores, college enrollment, use of social programs and Medicaid, and maternal care and stress. The data allow for a holistic view of policy impact; measuring gains to education and well-being from the immediate to the longer-term, and also measuring expenditure savings to government-funded social safety-net programs from early-life investments so that government can incorporate concepts of return on investment when considering how to get the most impact per dollar spent.

Our study finds that newborns just below the threshold who receive additional medical care fare significantly better later in life compared to those just above the threshold. Crossing the threshold is associated with increases in standardized test scores in elementary and middle school of 0.34 standard deviations, increases in college enrollment rates by 17.1 percentage points of a base rate of 53.6%, and decreases in social program expenditures of $27,291 by age 10 and $66,997 by age 14. Because the average cost of the additional medical services provided in the hospital at birth is around $4,000 [1], this study provides new facts to help policymakers evaluate the educational impact and potential financial returns of adjusting the threshold. We conclude that moving the threshold is a potential low-cost, high-impact policy lever for helping children at the margin to achieve better outcomes later in life.

To conduct this comprehensive study of outcomes for low birth weight newborns, we access data in *RI 360* that originate from several Rhode Island agencies. Three decades of birth records from the RI Department of Health define the study population of newborns with low birth weight. The RI Department of Education provides test scores from third, firth, and eighth grade standardized tests, the PSAT, the SAT, and Advanced Placement exams; records of grade repetition, Individualized Education Programs, and disciplinary actions; and college enrollment records from the National Student Clearinghouse. The RI Department of Human Services provides enrollment and benefit payment records for Supplemental Security Income, the Supplemental Nutrition Assistance Program, Medicaid, and Temporary Assistance for Needy Families. The RI Department of Labor and Training provides quarterly wage records that measure maternal employment rates and earnings following birth. The Centers for Disease Control provide survey responses from the Pregnancy Risk Assessment Monitoring System that measure maternal attitudes and experiences following birth.

Securing the data

Figure 1 summarizes our approach and highlights the first challenge when working with administrative records: deploying security controls that protect the data. Security is our first and foremost concern because the risks of improperly securing administrative data are great. Unauthorized access or data leakage have the potential for invasions of individual's privacy, identity theft, financial fraud, or even interference with our democratic institutions, including elections. Moreover, irresponsible handling of data can have spill-over effects which hinder scientific progress and policy improvement, as data owners perceive great risks of using data and partnering with scientists, even if the uses and partnerships are legitimate and secure.

We mitigate these risks by isolating all data ingest and processing within an encrypted tank (Figure 1a) inside a secure computing environment called a *data enclave* [16]. The enclave's key features are that it is physically secure and isolated from the Internet, data transfers in and out are restricted and subject to a documented approval process, all access is comprehensively audited, and access is granted to only a limited group of approved researchers. These security controls protect against unauthorized access and ensure that researchers access the data in compliance with the data sharing agreements governing their use.

Our implementation of the data enclave uses a locally-hosted system. However, modern cloud computing can help governments implement similar data enclaves using best practices for security and compliance. An additional benefit of a cloud solution is that government can own and operate the enclave, retain possession of the administrative data, and directly manage researchers' access, which removes the need for data transfers and data sharing agreements.

As an additional security measure, we restrict access to the encrypted tank using a two-party password, known only by senior leadership. A two-party password means two people each know a different half of the password, and both of the senior parties have to be present and consent to access the encrypted tank. This ensures that no individual researcher can access data that may reveal personally identifiable information.

Once the original data have been successfully transferred into the encrypted tank, we run an automated pipeline to separate out personally identifiable information (Figure 1b). Sensitive identification numbers – such as social security numbers or other identifiers deemed sensitive by the agency – are flagged ahead of time and automatically replaced with irreversible hashes, a technique that is widely used for protecting passwords [10]. Following this separation, the remaining data contain no personally identifiable information and are de-identified (Figure 1c).

Anonymizing the data

Once that data are secured, the next challenge is developing a method for identifying the same individual across data sets, while also preserving their anonymity so that researchers cannot discover their identity, even inadvertently. Although many of the data sources for the birth weight study identify records by social security number, an exception is the RI Department of Education, which identifies students by name and an internal identification number. Therefore,

we require an automated method to find matches among individual records based on hashed social security number when available, or else based on other fields like name and date of birth – all without revealing these fields to the researcher.

Our solution is to assign a global anonymous identifier (Figure 1d) to records right after separating out personally identifiable information. An automated script identifies matches among all hashed social security numbers, phonetic representations of names (using the Soundex algorithm [18]), and dates of birth. Using the global identifier, we can join information on outcomes to low birth weight newborns and their parents in the birth records without knowing any personally identifiable information for any of the individuals.

Our deterministic algorithm is designed to minimize false matches (incorrectly matching two different individuals) at the expense of having more missed-matches (in which two records of the same individual are not matched). Some records are missing too many fields and are considered too ambiguous to assign a global identifier, but this occurs for only 3.9% of records. As an alternative to the deterministic approach, the identifier could be constructed with probabilistic record-linkage methods that would likely have fewer missed-matches, but would also carry higher costs for computation and manual curation, as well as a higher likelihood of false-matches [12].

Integrating the data

We receive data extracts from administrative systems in various formats. The raw records used in the birth weight study arrive in the encrypted tank as comma-separated text (with varying delimiters and quoting conventions), fixed-width text, XML, and Excel files. Our approach has been to meet government data partners where they are, and to accommodate data extracts in the format they can most easily produce. Most agencies have perpetual operational demands on their administrative systems, and they are not resourced to support additional development for data warehousing or analytics.

Since there is no universal format or data dictionary across agencies, we normalize the data into a consistent format and typing structure with a lightweight and open-source integration tool called Secure Infrastructure for Research with Administrative Data. We developed this tool using an agile approach to meet the evolving needs of researchers and analysts as we built *RI 360*. Our GitHub repository[2] provides additional technical detail about our integration methods, as well as a worked example based on simulated data.

We chose an Extract Load Transform approach over the more typical Extract Transform Load approach [7]. In practice, this means that the de-identified data are loaded into *RI 360* in as close to their original format as possible. The majority of transformations are added later after researchers have a chance to perform preliminary analyses to assess data quality and understand the data-generating processes underlying the administrative systems.

[2] https://github.com/ripl-org/sirad-example

As an example, birth weight is an essential variable for defining our study population. However, is has been measured in different units (grams and ounces) over the three decades of birth records. Therefore, we construct a *birth derived* table that normalizes weight, as well as several other categorical variables measured at birth that switch from using numeric to character codes in the records over time. A derived table is a materialized view that aggregates, normalizes, and/or combines data from multiple original tables in *RI 360* into a single table that facilitates a specific analysis need – in this case, determining birth weight in a consistent way for all births. A more complex example in *RI 360* is the Supplemental Nutrition Assistance Program derived table. It combines records on applications, eligibility, benefit payments, and household structure to determine all individuals enrolled in the program at a given month and their household-level benefits.

At the highest level, we roll up all the derived tables into a single *RI 360* summary table, which spans 20 years of history for the state's most important programs and outcomes, as well as demographic information about anonymized individuals (e.g. age, race, ethnicity, and sex). Most of the outcomes in the birth weight study, including educational outcomes and benefit payments, are found in the *RI 360* summary table, which reduced the effort needed to launch the study. Creating derived tables also ensures that all studies using *RI 360* draw from common variable constructions and definitions that are robust and reproducible.

Supporting research integrity

A fundamental requirement of scientific findings is that they can be independently replicated by other investigators [17]. Similarly, fact-based policy should be based on robust findings that are peer-reviewed and replicable. To facilitate future replication, we update and snapshot *RI 360* approximately three times a year, creating what we call a *research version* (Figure 1f). The research versions are de-identified data and become the permanent archive of *RI 360*. We have currently generated 11 such versions. Once a research version has been validated, the encrypted original data used to create that version are wiped from the encrypted tank and destroyed. Every analysis is tied to a fixed research version of the database, and can be rerun against the research version at a later time to replicate the results. Additionally, to encourage reproducibility, analysis projects use a common project template to organize code and research results in a standardized way.[3]

Even through *RI 360* has been de-identified, our data sharing agreements restrict all research with anonymized individual-level records to the data enclave. Only aggregated or statistical results such as summary tables, plots, and regression coefficients can be exported from the enclave. All statistics must be aggregated such that they represent 11 or more distinct individuals. To ensure compliance with usage agreements and security, no individual researcher has the ability to export files from the enclave. Copy and paste functionality has also been disabled within the enclave's user interface. Exports are subject to review and documentation to ensure that exported results conform to usage agreements (Figure 1g), and they trigger real-

[3] https://github.com/ripl-org/predictive-template

time alerting to senior leadership. A read-only snapshot of each export is archived in the enclave to facilitate future audits.

Conclusion

The insights gained from research with administrative data have the potential to transform the way that policymakers approach some of society's most important policy decisions. Robust evidence on previous policy outcomes and predictive modeling of future outcomes can guide policymakers to smarter policies with greater benefits at lower cost. We have described a comprehensive approach to overcoming the many challenges faced when integrating siloed state-wide databases into a data infrastructure for fact-based policy, which is the first system of its kinds in the U.S. In the future, we hope that more systems of this kind will provide policymakers at all levels of government, and in many countries across the world, with a rich ecosystem and evidence base for the important decisions they make on behalf of their constituents.

Acknowledgments

This work was supported by the Smith Richardson Foundation and Laura & John Arnold Foundation.

References

1. Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. 2010. Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborn. *Quarterly Journal of Economics* 125, 2 (May 2010), 591–634. DOI: https://doi.org/10.1162/qjec.2010.125.2.591
2. Raj Chetty, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA* 315, 16 (April 2016), 1750–1766. DOI:https://doi.org/10.1001/jama.2016.4226
3. Eric Chyn, Samantha Gold, Justine S. Hastings. (Forthcoming). Short- and long-run impacts of health interventions for very low birth weight children.
4. Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking. Retrieved from https://www.cep.gov/cep-final-report.html.
5. Roxanne Connelly, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research* 59, (September 2016), 1–12. DOI:https://doi.org/10.1016/j.ssresearch.2016.04.015

6. Jonathan M.V. Davis and Sara Heller. 2017. *Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs*. Working Paper No. 23443. National Bureau of Economic Research, Cambridge, MA. DOI:https://doi.org/10.3386/w23443

7. Umeshwar Dayal, Malu Castellanos, Alkis Simitsis, and Kevin Wilkinson. 2009. Data integration flows for business intelligence. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology,* Saint Petersburg, Russia, March 24-26, 2009. DOI:https://doi.org/10.1145/1516360.1516362

8. William H. Dutton and Kenneth L. Kraemer. 1985. *Modeling as Negotiating: The Political Dynamics of Computer Models in the Policy Process*. Ablex Publishing Corporation, Norwood, NJ.

9. James Danziger. 1977. Computers and the Frustrated Chief Executive. *Management Information Systems Quarterly* 1, 2 (June 1977), 43–53.

10. Liran Einav and Jonathan Levin. 2014. Economics in the age of big data. *Science* 346, 6210 (2014), 1243089.

11. P. Gauravaram. 2012. Security Analysis of salt||password Hashes. In *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, 25–30. DOI:https://doi.org/10.1109/ACSAT.2012.49

12. Katie Harron, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L Barreto, and Harvey Goldstein. 2017. Challenges in administrative data linkage for research. *Big Data & Society* 4, 2 (December 2017), 2053951717745678. DOI:https://doi.org/10.1177/2053951717745678

13. Justine S Hastings. 2019. Fact-Based Policy: How Do State and Local Governments Accomplish It? The Hamilton Project (Brookings Institution), Policy Proposal 2019-01. Retrieved from http://www.hamiltonproject.org/assets/files/Hastings_PP_web_20190128.pdf

14. Justine S. Hastings and Jesse M Shapiro. 2017. How Are SNAP Benefits Spent? Working Paper No. 23112. National Bureau of Economic Research, Cambridge, MA. Retrieved from http://www.nber.org/papers/w23112.

15. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *Q J Econ* 133, 1 (February 2018), 237–293. DOI:https://doi.org/10.1093/qje/qjx032

16. Julia Lane and Stephanie Shipp. 2007. Using a Remote Access Data Enclave for Data Dissemination. *International Journal of Digital Curation* 2, 1 (2007), 128–134. DOI:https://doi.org/10.2218/ijdc.v2i1.20

17. Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (December 2011), 1226–1227. DOI:https://doi.org/10.1126/science.1213847

18. C. Russell Robert. 1918. The Soundex coding system. Patent No. US1261167.

Figure 1. Overview of the processing steps to secure, integrate, and conduct anonymized research with administrative data. Agencies securely transfer data extracts to an encrypted tank inside the data enclave **(a)**. These data are split **(b)** into personally identifiable information and de-identified data **(c)**. Personally identifiable information is used to construct an anonymized global identified **(d)** and to geocode home addresses to construct an anonymized neighborhood identifier **(e)**. De-identified data are used to construct research versions of the *RI 360* database **(f)**, which can be accessed by approved researchers from inside the data enclave. Research findings can be exported from the data enclave through a documented review process **(g)**.