BIG DATA AND THE TRANSFORMATION OF PUBLIC POLICY ANALYSIS

Ron S. Jarmin and Amy B. O'Hara

INTRODUCTION

Recent years have seen a growing number of uses of novel "big data" sources to monitor, improve, and study the delivery of public services. By "big data," we mean data generated as a consequence of government, business, or citizen activities. Big data is often said to be characterized by four Vs: volume, velocity, variety, and veracity. Examples include data from cameras and sensors that permit real-time traffic monitoring to more novel uses such as tracking gunshots through networks of auditory sensors. The latter allow more granular measurement of criminal activity than traditional police reports (Carr & Doleac, 2015) and are suggestive of how valuable these new data sources can be to policy analysts and applied social scientists.

New data from sensors or apps where citizens engage directly with government service providers (Crawford & Walters, 2013) will undoubtedly help transform public policy analysis. However, many programs directly impact individuals, households, firms, or other private sector organizations. These programs generate their own administrative "big data." Indeed, the administrative records these programs produce often have volume, an assembly of them provides variety, and depending on source, they have veracity (UNECE, 2013). Improvements in computation and analytics have streamlined processing of such data, allowing analysts to draw actionable intelligence from them. Said differently, compiling rich data resources, especially when linked and analyzed with modern "big data" tools, provides unprecedented opportunity to transform public policy analysis.

However, administrative data have varying privacy or confidentiality statutes or regulations making access difficult. To address this, Congress recently approved funding for the Census Bureau to fortify its platform aligning the appropriate data, tools, and researchers to facilitate evidence building. Relevant program data is held in a variety of federal and state agencies. These entities possess differing capabilities to support research, have data that is not prepared for analytical use, and have a variety of access requirements and processes. A federated infrastructure permits data from many sources to be curated, integrated, and provisioned in ways that foster credible and transparent evidence building and adhere to the rules and policies of the relevant data owners. The platform being expanded at the Census Bureau addresses the key barriers to research access while permitting information providers to retain control of their data. Housing the infrastructure at the Census Bureau makes sense as it has broad legal authority to obtain data, a robust infrastructure for curating and integrating data from many sources, and an outstanding record as a data steward. This approach is efficient, consistent with privacy principles, and clarifies access procedures by providing one "front door" for policy analysts and researchers needing access to administrative data.

Improving access to administrative records and technology to link and analyze these "big data" will transform evidence building. As we highlight below, this change is already underway, with analyses that investigate the impacts of policies over

Journal of Policy Analysis and Management, Vol. 00, No. 0, 1–7 (2016)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21925

time, across programs, and spanning generations. Moreover, the Census Bureau has used administrative records to measure economic activity since the 1940s and is currently adopting best practices from the private sector to modernize economic measurement (Bostic, Jarmin, & Moyer, 2016). Thus, our big data solution for policy analysis is a straightforward extension of these activities.

For social scientists and policy analysts, however, big data sources can complement data carefully "designed" for a given purpose. Surveys are designed data: offering a known universe/frame, stable and well-defined questions and content, and documented treatment of missing data. These have been widely used to analyze the impact of policy changes and longitudinal outcomes using weighted samples. Surveys have a lot of variety, but they fall far short on volume, velocity and, increasingly, on veracity. Surveys have quality challenges as a result of declining respondent cooperation and participation (Meyer, Mok, & Sullivan, 2015). Studies show that underreporting in surveys (Call et al., 2013; Meyer & Mittag, 2015) impacts understanding of welfare and Medicaid participation, using administrative data to show the measurement differences. Using the sources together holds great promise to produce blended statistics. For example, blended statistics could help income studies beset by right tail problems such as top-coding, undercoverage, and underreporting (Burkhauser et al., 2012). Income studies have turned to administrative data (Auten, Gee, & Turner, 2013; Chetty, Hendren, & Katz, 2014) despite limited information on the left tail. Surveys have traditionally offered the only view of individuals over time. Longitudinal surveys such as the National Longitudinal Survey, Panel Study on Income Dynamics, and National Center for Education Statistics surveys that have been mainstays in education evaluations are giving way to studies relying on administrative data (Figlio, Karbownik, & Salvanes, 2015).

ILLUSTRATIVE USES OF ADMINISTRATIVE DATA FOR POLICY ANALYSIS

In this section, we discuss several recent papers as "isolated uses" of using a variety of administrative data sets.[1] These studies are typically conducted by teams that obtained access to confidential data through arrangements with administrative or statistical agencies or through collaboration with agency researchers. We argue that this narrowly defined data access, underdeveloped data infrastructure, and inadequate metadata and tools prevent all but highly expert researchers to draw inferences from the data. We will return to access issues in the next section but want to illustrate more specifically how "big data" can aid in policy analysis.

Intergenerational mobility studies have been transformed using longitudinal tax data. The Joint Statistical Research Program of the Statistics of Income Division at the IRS enables studies that use long panels of tax returns to observe individuals over time. Isolated access[2] for specific tax administration research resulted in the Equality of Opportunity project,[3] deemed "big data" by lead researcher Raj Chetty. Through that data access, Chetty, Hendren, and Katz (2014) analyzed geography and high-mobility areas. Also using tax data and collaborating with IRS analysts, a Stanford team has studied intergenerational mobility (Mitnik et al., 2015). Johnson,

[1] While not our focus here, the recent empirical literature features studies using an increasing number and variety of private sector data sets. The potential of these data resources for public policy analysis is only beginning to be explored. Einav and Levin (2014a, 2014b) review recent work using private data. Researchers affiliated with NSF-Census Research Network have been exploring the use of Twitter data to understand labor market transitions and personal financial management tools to understand household income and spending patterns (see Gelman et al., 2014).

[2] See http://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records.

[3] See http://www.equality-of-opportunity.org/.

Massey, and O'Hara (2015) discuss the potential and challenges of using tax data, survey data, and other administrative data to study mobility.

Chetty, Hendren, and Katz's (2016) follow-up on the Moving to Opportunity (MTO) experiment assesses long-term outcome for the voucher program that started in 1994. Earlier studies failed to find impacts on earnings or employment rates of adults and older children from MTO treatments (e.g., Kling, Liebman, & Katz, 2007; Oreopoulos, 2003; Sanbonmatsu et al., 2006). However, using tax returns from 1996 to 2012, Chetty, Hendren, and Katz (2016) found that moves to lower poverty neighborhoods significantly improved college attendance rates and earnings for children who were young (below age 13) when their families moved. In his summary of MTO studies, Rothwell (2015) notes, "As Chetty and his colleagues show, even a few extra years of data can make a large difference." To conduct the study, Chetty, Hendren, and Katz (2016) had isolated access to the MTO data as well as tax return data.

Our final example of isolated access is led by the University of Michigan's Institute for Research on Science and Innovation (IRIS).[4] IRIS is a consortium of academic institutions that is leveraging and linking its administrative data on federal grant expenditures to analyze its research portfolios, document its varied economic impact, and generate new scholarly findings. To do this, detailed data on federally funded research grants at universities is linked to Census Bureau survey, census, and administrative data assets at the micro-level. The research team utilizes modern "big data" tools such as machine learning and network analysis to produce new statistics on the research value chain, the dissemination of scientific knowledge and expertise, and links to innovation, economic growth, job creation, and entrepreneurship.

The IRIS project illustrates how data and linkage infrastructure, combined with modern analytics, provide policy analysts with unprecedented capabilities for measuring impacts. The infrastructure has produced studies on the job placements (especially for students and graduate students) and entrepreneurial activity of university-funded researchers (Zolas et al., 2015), and the impact of university research on regional economies (Goldschlag et al., 2016).

The mobility, MTO, and IRIS analyses would be impossible without accessing administrative records data. Linking these rich data sources at the microlevel vastly increases the depth and breadth of feasible policy analysis. In the next section, we describe Census Bureau's efforts to expand such data access.

EXPANDING ACCESS TO ADMINISTRATIVE RECORDS

While administrative records are not synonymous with big data, the mobility, MTO, and IRIS projects demonstrate the importance of large, curated, and accessible federal and state agency files. The Census Bureau has a long history of using administrative records to improve data quality, control data collection costs, reduce respondent burden, and develop new data products. The Census Bureau's uses of administrative records and record linkage are grounded in strong laws that guide how the Census Bureau both accesses and protects administrative records.

In FY 2016, the Census Bureau received $10 million to accelerate the expansion of its administrative records infrastructure.[5] The President's proposed budget specified that this infrastructure investment should have four key goals: (1) expedite the acquisition of federal and federally sponsored administrative data sources; (2) improve data documentation and linkage techniques; (3) leverage and extend existing

[4] See http://iris.isr.umich.edu/.
[5] See https://www.whitehouse.gov/sites/default/files/omb/budget/fy2016/assets/ap_7_evidence.pdf.

systems for governance, privacy protection, and secure access to this data; and (4) administer a Commission on Evidence-based Policymaking.

The Census Bureau has significant expertise and infrastructure dedicated to acquiring federal and state program data. Census obtains records from the Social Security Administration, Housing and Urban Development, Indian Health Service, and the Centers for Medicare and Medicaid Services. It also obtains state data for programs such as the Unemployment Insurance, Supplemental Nutritional Assistance Program, and Women Infants and Children Program. Through our infrastructure initiative, we will strive to obtain new federal and state data sources. The Census Bureau will work with the Office of Management and Budget (OMB), government, social science, and policy stakeholders to develop a list of high-priority files to pursue and streamline procedures in order to add new data to the infrastructure.

A stable and robust data infrastructure will contain uniformly processed, documented data to observe trends within and across states and over time. The Census Bureau is partnering with vendors and academic experts to develop methods and software to improve data harmonization. Whenever permitted by the providing agency, Census will develop and implement greater automation to document, link, and curate the sources. As shown in other countries, coordination with the data producer improves metadata. The Census Bureau will explore ways to broaden administrative or legislative access and to encourage standardization of data formats/interoperability.

The Census Bureau will provide data access procedures through two mechanisms. First, in conjunction with data-providing agencies, the Census will expand the technical infrastructure of the Federal Statistical Research Data Centers[6] (FSRDCs) and develop straightforward procedures for accessing the data in those facilities. Second, the Census Bureau will document the procedures by which it can provide linkage services for other agencies or entities. This would allow the use of uniformly processed and linked files for use at the source agency, in the FSRDC network, and potentially through other secured computing platforms approved by the data-providing agencies.

The accelerated development of acquisition, data processing, and data access will be supported by a Commission on Evidence-based Policymaking. The Commission will comprise academic researchers, data experts, and those with experience administering programs. The Commission will determine how data should be accessed to facilitate program evaluation, continuous improvement, policy-relevant research, and cost-benefit analyses by qualified researchers and institutions. The Commission will consider how a clearing house could be self-funded; which types of researchers, officials, and institutions should have access to data and what their qualifications should be; what limitations should be placed on the use of data provided; how to protect information and ensure individual privacy and confidentiality; and incentives to facilitate interagency information sharing.

SUSTAINING THE INFRASTRUCTURE

While the Census Bureau infrastructure is in early stages, we must consider how to sustain the effort. A successful infrastructure will be accessible, fair, and open. It needs to be planned to allow growth in data users and data sources. Its access protocols must be fair and meet ethical standards (Fantuzzo & Culhane, 2015). The purpose of the infrastructure must be clear and openly described.

[6] More information on the FSRDC network is available at http://www.census.gov/fsrdc.

The infrastructure will exist through trust in the Census Bureau's data stewardship and technical controls, and respect for privacy and confidentiality. Title 13 of the U.S. Code provides an explicit legal framework for the Census Bureau to acquire, use, and protect administrative records. The privacy of data from administrative records is held to the same standard as data collected from census and survey respondents. The Privacy Act similarly provides for both protection of individual records as well as a framework permitting their use for statistical purposes. The Act prohibits agencies from incorporating individual identifiers into a "system of records," without the prior written consent of the individual to whom the information pertains. However, the statute provides a limited number of exceptions to the requirement of prior consent, including an exception for the Census Bureau for activities pursuant to Title 13.

We need to demonstrate that honest statistics can be generated using big, administrative, or blended data sources. The infrastructure must support data quality and validation work, drawing from the experiences of researchers and evaluators to improve the data, to build the codebase for evaluation. This crowdsourcing approach will also be essential as analyses take advantage of emerging data sources. We will need to continue conducting surveys to see if new and evolving data sources make sense. This approach is consistent with the big data explorations happening around the world in National Statistical Institutes (NSIs), including the need for calibration studies using survey data (Dunne, 2013). We must explore the changing data landscape—and its inherent variability—and communicate its potential and challenges to end users. More work is needed to produce safe outputs that protect individuals from reidentification and to explore synthetic data. We need to sustain an environment that increases data access while maintaining confidentiality.

Another key to sustaining the infrastructure is investing in human capital to use the data well. Not only are the data different and evolving, the types and tools for analysis are changing. We need to put the data science tool kit into the hands of social scientists. The research and evaluation community needs to hire, train, and retain the evidence builders. This requires investment in internships, fellowships, postdocs, and agency rotations.

CONCLUSION

A Census Bureau data infrastructure is the best path forward, leveraging statutory authority to access data and use it for statistical purposes such as program evaluation, building upon its linkage infrastructure, and relying on its public trust and strong data stewardship. The agency coordinates within the Federal Statistical System to encourage data linkages that build evidence. The infrastructure enables long-term follow-up on program participants or cohorts like the Rhodes et al. (2014) study using administrative records showing that most released offenders never return to prison. The infrastructure will reduce the frictions and costs of data access to produce near real-time analyses to help programs avoid costs or negative outcomes, facilitating more studies of individuals with high utilization of health services and law enforcement systems, so-called frequent fliers (Ryan, Hendler, & Bennett, 2015). The infrastructure will support studies that improve policy administration, offering a secure facility to integrate files and apply big data tools to build evidence.

These efforts align with bipartisan legislation, administration objectives, and years of preparation in the Federal Statistical System. The Census Bureau has engaged in previous demonstrations and quasi-experimental analyses linking administrative program data for statistical purposes (e.g., Jarmin, 1999; Tangka et al., 2015; Brown & Earle, 2012), and looks forward to supporting this work. Importantly, these efforts

not only improve data for evidence building and policy analysis, but also strengthen the data infrastructure that underlies federal statistics.

RON S. JARMIN is the Assistant Director for Research and Methodology at the U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 (e-mail: ron.s.jarmin@census.gov).

AMY B. O'HARA is the Chief of the Center for Administrative Records Research and Applications at the U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 (e-mail: amy.b.ohara@census.gov).

ACKNOWLEDGMENTS

The conclusions and opinions expressed here are those of the authors and do not necessarily reflect the views of the U.S. Census Bureau.

REFERENCES

Auten, G., Gee, G., & Turner, N. (2013). Income inequality, mobility, and turnover at the top in the U.S., 1987–2010. American Economic Review, 103, 168–172.

Bostic, W. G., Jarmin, R. S., & Moyer, B. (2016). Modernizing federal economic statistics. American Economic Review, 106(5).

Brown, J. D., & Earle, J. S. (2012). Do SBA loans create jobs? Estimates from universal panel data and longitudinal matching methods. SSRN Electronic Journal. Retrieved January 31, 2014, from http://doi.org/10.2139/ssrn.2205174.

Burkhauser, R. V., Feng, S., Jenkins, S. P., & Larrimore, J. (2012). Recent trends in top income shares in the United States: Reconciling estimates from March CPS and IRS tax return data. Review of Economics and Statistics, 94, 371–388.

Call, K. T., Davern, M. E., Klerman, J. A., & Lynch, V. (2013). Comparing errors in Medicaid reporting across surveys: Evidence to date. Health Services Research, 48, 652–664.

Carr, J., & Doleac, J. (2015). The geography, incidence, and underreporting of gun violence: New evidence using ShotSpotter data. Retrieved April 28, 2016, from http://jenniferdoleac.com/wp-content/uploads/2015/03/Carr_Doleac_gunfire_underreporting.pdf.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. Quarterly Journal of Economics, 129, 1553–1623.

Chetty, R., Hendren, N., & Katz, L. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity project. American Economic Review, 106(4), 855–902.

Crawford, S. P., & Walters, D. (2013). Citizen-centered governance: The mayor's office of new urban mechanics and the evolution of CRM in Boston. SSRN Electronic Journal. Retrieved April 28, 2016, from http://doi.org/10.2139/ssrn.2307158.

Dunne, J. (2013). Big data coming soon . . . to an NSI near you. Proceedings 59th ISI World Statistics Congress. Retrieved from http://2013.isiproceedings.org/Files/STS018-P3-S.pdf.

Einav, L., & Levin, J. (2014a). The data revolution and economic analysis. Innovation Policy and the Economy, 14, 1–24.

Einav, L., & Levin, J. (2014b). Economics in the age of big data. Science, 346, 1243089–243094.

Fantuzzo, J., & Culhane, D. (Eds.). (2015). Actionable intelligence—Using integrated data systems to achieve a more effective, efficient, and ethical government. New York: Palgrave Macmillan.

Figlio, D. N., Karbownik, K., & Salvanes, K. G. (2015). Education research and administrative data (no. w21592). Retrieved February 15, 2016, from http://www.nber.org/papers/w21592.

Gelman, M., Kariv, S., Shapiro, M. D., Silverman, D., & Tadelis, S. (2014). Harnessing naturally occurring data to measure the response of spending to income. Science, 345, 212–215.

Goldschlag, N., Bianchini, S., Jarmin, R., Llerena, P., McFafdden-Allen, B., Owen-Smith, J., SanMartin Sola, J., Weinberg, B., Zolas, N., & Lane, J. (2016). Research Funding and Regional Economies. Working Paper.

Jarmin, R. S. (1999). Evaluating the impact of manufacturing extension on productivity growth. Journal of Policy Analysis and Management, 18, 99–119.

Johnson, D., Massey, C., & O'Hara, A. (2015). The opportunities and challenges of using administrative data linkages to evaluate mobility. Annals of the American Academy of Political Science, 657, 247–264.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. Econometrica, 75, 83–119.

Meyer, B. D., & Mittag, N. (2015). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net (no. w21676). Retrieved October 26, 2015, from http://www.nber.org/papers/w21676.

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household surveys in crisis. Journal of Economic Perspectives, 29, 199–226. Retrieved November 6, 2015, from http://www.aeaweb.org/articles.php?doi=10.1257/jep.29.4.199.

Mitnik, P., Bryant, V., Weber, M., & Grusky, D. (2015). New estimates of intergenerational mobility using administrative data. Retrieved February 16, 2016, from https://www.irs.gov/pub/irs-soi/15rpintergenmobility.pdf.

Oreopoulos, P. (2003). The long-run consequences of living in a poor neighborhood. Quarterly Journal of Economics, 118, 1533–1575.

Rhodes, W., Gaes, G., Luallen, J., Kling, R., Rich, T., & Shively, M. (2014). Following incarceration, most released offenders never return to prison. Crime & Delinquency. Retrieved December 8, 2015, fromhttp://doi.org/10.1177/0011128714549655.

Rothwell, J. (2015). Sociology's revenge: Moving to opportunity (MTO) revisited. Retrieved February 16, 2016, from http://www.brookings.edu/blogs/social-mobility-memos/posts/2015/05/06-moving-to-opportunity-revisited-rothwell.

Ryan, J., Hendler, J., & Bennett, K. P. (2015). Understanding emergency department 72-hour revisits among medicaid patients using electronic healthcare records. Big Data, 3, 238–248.

Sanbonmatsu, L., Kling, J. R., Duncan, G. J., & Brooks-Gunn, J. (2006). Neighborhoods and academic achievement: Results from the moving to opportunity experiment. Journal of Human Resources, XLI, 649–691.

Tangka, F. K. L., Howard, D. H., Royalty, J., Dalzell, L. P., Miller, J., O'Hara, B. J., Sabatino, S. A., Joseph, K., Kenney, K., Guy, G. P., & Hall, I. J. (2015). Cervical cancer screening of underserved women in the United States: Results from the National Breast and Cervical Cancer Early Detection Program, 1997–2012. Cancer Causes & Control: CCC, 26, 671–686.

UNECE. (2013). What does big data mean for official statistics? Retrieved February 2, 2015, from http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Smith, J. O., Rosen, R. F., Allen, B. M., Weinberg, B. A., & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for Ph.D. recipients. Science, 350, 1367–1371.