Using Administrative Big Data to Solve Problems in Social Science and Policy Research*

Xi Song

Department of Sociology
University of Pennsylvania

Thomas S. Coleman

Harris School of Public Policy
University of Chicago

November 2020

* This article was translated into Korean and published in *Global Social Security Review*. Please cite Song, Xi and Thomas S. Coleman. forthcoming. "Using Administrative Big Data to Solve Problems in Social Science and Policy Research." *Global Social Security Review* 14: 5–15. Send correspondence to Xi Song, Department of Sociology, University of Pennsylvania, 3718 Locust Walk, Philadelphia, Pennsylvania, 19104 (xisong@upenn.edu).

Abstract

This article describes an explosion in the availability of individual-level public administrative data in the United States and worldwide. These datasets can be used as stand-alone resources or linked across different sources. These new resources will facilitate transformative research on social, demographic, and economic changes, policy evaluation, and other experimental analyses. We discuss the current status of administrative big data in the United States, their potential to advance social science and policy studies, and advantages and challenges for using these data in practice. We showcase a few ongoing large-scale U.S. administrative data initiatives and hope to spark future parallel endeavors in other countries.

Introduction

Big data—the rapid growth in data volume, variety, and velocity—are revolutionizing social, behavioral, and policy research. A vast amount of data emerges from social media, online registrations, transactions, record keeping, satellite and GPS tracking devices, natural languages, and information networks. Digital archives have grown exponentially, thanks to automation, machine learning, and information technology. New forms of data, such as text, pictures, videos, geolocations, have broadened the concepts of data in social science research. The rise of crowdsourcing platforms makes data collection easier, cheaper, and faster. Naturally occurring, real-time data minimizes the role of researchers in the data-generating process. We focus on administrative microdata, one form of big data. Social science researchers have benefited and will continue to benefit from administrative personal records—census records, tax return files, vital statistics, parish records, voter registration, medical claims, family genealogies, and population registers—and the integration of such microdata across disparate administrative and other data sources.

Empirical research in the social sciences has traditionally relied on survey data, intentionally designed and collected for research or policy purposes. In contrast, today's world is awash in "Big Data" from both administrative and private-sector data sources. This article provides an overview of the current policy and practice of using public administrative data in social science research, especially in the context of the United States. We first contrast survey data versus administrative data, then discuss benefits and cautions of using and linking across administrative data for academic research and policy analyses. We then provide examples of ongoing data linkage projects in the United States with brief discussions about their background, achievement, and challenges in privacy protection and

implementation issues. We conclude by contrasting with other forms of big data and considering future developments in the U.S. and implications for similar endeavors in other countries.

Survey Data

There are no widely accepted definitions of survey data. In their classic book on survey methodology, Groves et al. (2009) consider censuses as the earliest type of survey, whereas in the private sector, the terms of surveys and polls are typically used interchangeably. For the purpose of this paper, we define surveys as data that are collected for a subset of a population, with the intent to provide information on the overall population (Slemrod 2016). Surveys are generally designed to elicit information on some particular underlying construct or relationship of interest to the researcher or agency—there is thought and effort expended to connect the survey data with relevant theoretical constructs. Surveys are subject to four primary sources of error: (1) coverage error, (2) nonresponse error, (3) sampling error, and (4) survey measurement error (inaccuracies in responses). Survey data may be numeric or text and may be collected in any variety of forms, from face-to-face or mailed questionnaires in the early days of survey research (e.g., in the 1930s) to fully electronic questionnaires today (Groves 2004). Compared to administrative data assembled and maintained exclusively by official government agencies, survey data can be collected by public or private institutions for registration, research, marketing, or political prediction purposes.

Administrative Data

Administrative data, in contrast to survey data, are generally assembled and maintained by

official government agencies as part of managing and administering programs, such as vital registration, census records, tax collection (the IRS), state unemployment insurance, retirement (Social Security Administration wage and salary records), medical coverage (the Centers for Medicare and Medicaid), schooling, and education. Administrative data have three important benefits relative to survey data (as highlighted by Card et al. 2010):

1. Large size: usually covering the whole population rather than a sample or subset, reducing sampling error and allowing the study of rare events;
2. High-quality information: responses generally have consequences for respondents, either positive as in the paying of benefits (for example with unemployment insurance benefits) or negative as in penalties for nonresponse or false response (fines and jail time for falsifying tax returns);
3. Longitudinal: usually covering the same units over time, providing a natural panel structure.

Research Benefits of Administrative Data

Administrative data provide a variety of benefits for social science researchers. The large sample size opens opportunities to study small subgroups, rare events, and groups that might not be captured in existing surveys. Lower nonresponse rates and high-quality data allow re-examining accepted bodies of work and the development of new insights. The inclusion of multiple cohorts enables the study of longitudinal process and social change over time. Finally, administrative data can offer savings in the collection of primary data.

The study of top income shares and inequality using U.S. tax data, pioneered by Piketty and Saez (2003), exemplifies the benefits of large sample and (synthetic) cohorts. The top 1 percent or top 0.1 percent of income earners are typically invisible in standard

social surveys, for two reasons. First, survey samples usually target the general population and are thus too small to make reliable inferences for the earnings of populations at the extremes. Second, to protect respondents' identities in the relatively small surveys, incomes are often top-coded (for example, any income higher than $200,000 is reported as "greater than $200k" rather than the actual amount). Access to Internal Revenue Service (IRS) administrative data and the full universe of tax returns has allowed the study of the top 1 percent and top 0.1 percent of the income distribution. These data have been extended back in history over roughly 100 years, challenging our understanding of both the current level and the history of inequality in the U.S.

Meyer and Mittag (2019) provides an example of using administrative data to reexamine and improve poverty measures. They find that improved measurement substantially alters the apparent effectiveness of government poverty programs. The Current Population Survey (CPS), specifically the March supplement (Annual Social and Economic Supplement or ASEC), is the standard survey source for individual and household income. Specifically, they merged data from the CPS, New York social service agencies, and the Federal Department of Housing and Urban Development (HUD) to compare survey-reported and administrative payments for government cash transfer programs. Amounts reported in the CPS for SNAP (food stamps), TANF (Temporary Assistance for Needy Families), General Assistance, and federal housing assistance are substantially lower than actually paid. Proportionally, this discrepancy has the largest effect at the lower end of the income distribution, enough to substantially change our understanding of the fraction of the population in poverty. This work has helped inform the policy debate over producing additional measures of poverty (see OMB 2020).

Research Challenges of Administrative Data

Administrative data may go far in reducing or eliminating the survey-data errors mentioned above, but they do not solve all survey-data, or research, problems. Administrative data may contain measurement errors, reporting errors, problems with record matching and in constructing the statistical units of interest, or problems because the particular administrative measure differs from what an analyst requires (Groen 2012). It is important to understand and document the data definition and generation process—administrative data are usually designed for a specific administrative purpose and do not *necessarily* match the constructs researchers want to study (Connelly et al. 2016). Goerge and Lee (2002) provides an overview of the matching and cleaning necessary when using administrative data.

The IRS tax data discussed above, used in the study of top income shares, provides a good example of the care that must be paid to data definition. One would assume that income reported on tax returns provides a simple and clean measure of income. However, tax data are far from perfect. Slemrod (2016) lays out many of the issues, among which two are of particular relevance for comparisons across time (also see Auten and Splinter 2019 and Guvenen and Kaplan 2017). First, changes in tax rules may change how individuals report income with no change in actual income. The Tax Reform Act of 1986 changed statutory rates for income reported as personal versus corporate income and thus changed incentives for reporting income as personal income on individual returns, even with no change in underlying income. Differentially changing incentives at the top and bottom of the distribution introduce complications into comparisons across time. Second, tax data

also provide an example of issues concerning the unit of measurement. While measuring income by tax return seems a natural solution and is often made in academic work, policy concerns focus on individuals or households, not tax returns. Changing demographics—marriage rates in particular—can change the unit of measurement, again differently at the top and bottom of the income distribution. Such changes can have substantive effects on measured top income shares. Both of these issues can be managed, but researchers must be open to recognizing and addressing them.

One important challenge using administrative data is the frequent lack of ancillary data, such as demographic data for tax filers. Linking administrative data across datasets can go far toward addressing the first issue. Many northern European countries (Norway, Finland, Sweden, Denmark stand out) have built such linked datasets (see United Nations 2007) and are ahead of the U.S. in developing comprehensive linked datasets. We discuss data linkage issues in the next section.

Linking Administrative and Survey Data

In Philip Smith and Barbara Boyle Torrey's prescient *Science* paper of 1996, "The Future of the Behavioral and Social Sciences," the first challenge that they presented was "to integrate current data sets." A centralized agency that compiles, integrates, and provides access to administrative data is a reasonable model for many countries—Statistics Denmark is one example. For the United States this model is less attractive, for several reasons (as discussed in Card et al. 2010). First, the US government is decentralized, with multiple agencies at the three levels of federal, state, and local government. Second, different agencies are covered by different laws for sharing and privacy, and centralization would

require potentially difficult legislative action. Third, the US has a long tradition of distrust of centralized government and particularly concentration of power in a single agency, and development of administrative data sharing must, of necessity, respect this tradition. One result of the decentralized US approach is that privacy concerns are handled on a more ad-hoc basis rather than with a centralized communication strategy.

In the United States, more than 70 federal government agencies, and an even greater number of private agencies, have collected survey and administrative data. However, each survey is often collected for its own purpose and a particular targeted population. Most of these data sources are not directly comparable or connected. From an academic perspective, researchers wish to re-use the huge amount of microdata collected by the government and third-party institutions through record linkage and statistical matching to address important research problems. From a governance perspective, public policymakers and government officials are under increasing pressure to evaluate their programs and improve policy designs and decisions.

Private research and federal institutions in the U.S. have started to create public databases with consistent formats that link different administrative data sources and other social surveys. Among these research efforts, the American Opportunity Study (AOS), spearheaded by a group of sociologists and economists commissioned by the National Research Council (NRC) and the Census Bureau, has linked the censuses of 1960 through 2010 and the American Community Survey (ACS) (Grusky et al. 2019). The Longitudinal, Intergenerational Family Electronic Micro-Database Project (LIFE-M) pioneered by a group of social scientists at the University of Michigan has merged birth, death, and marriage records with the 1940 Census (Bailey et al. 2019). The Minnesota Population

Center Integrated Public Use Micro-data Series (IPUMS) has collected and distributed a vast array of population data from around the world (Ruggles et al. 2015). Linking these datasets has substantially expanded the number of variables and time coverage of the sample, yielding more reliable data for statistical analyses.

At the Federal level and as a leading effort to address these needs, the U.S. Census Bureau has established the Center for Administrative Records Research and Applications (CARRA), which undertakes the mission of assessing census and survey data collection operations, improving data quality and safety, and combining information from multiple data sources to generate new products for research and policy-use purposes that cannot be achieved through single data sets. CARRA has developed matching software to link person records across datasets following the long tradition of data dissemination adopted by governments in many other European countries. Below we illustrate linking procedures and data privacy issues using the Census Longitudinal Infrastructure Project (CLIP).

The Census Longitudinal Infrastructure Project aims to create a set of linked data files from decennial censuses, surveys, and administrative records collected by the Census Bureau.[1] The linkage system relies on Protected Identification Keys (PIK), which were assigned by the Person Identification Validation System (PVS) as a probabilistic matching algorithm that compares characteristics of the Census Bureau's records to characteristics of records in a reference file constructed from Social Security Administration's NUMIDENT file and other administrative data. These variables include Social Security Numbers (SSN), full name, date of birth, address, and parents' names, depending on the available information in the data. Each PIK uniquely identifies a particular person, thus allowing

[1] https://www.census.gov/about/adrm/linkage/projects/clip.html

researchers to find individuals across multiple PIKed data sources. Ferrie, Massey, and Rothbaum (2016) and Massey et al. (2018) have systematically documented details on the data linkage using PIKs based on PVS, linked sample representativeness, and potential biases.

An on-going effort at the U.S. Census Bureau is the development of the Comprehensive Income Dataset (CID), a prototype for a restricted micro-level dataset that combines the demographic detail of survey data with the accuracy of administrative measures (Medalia at al. 2019). The CID incorporates information on nearly all taxable income, tax credits, and cash and in-kind government transfers, with the goal of providing an accurate and comprehensive measure of income for the population of United States individuals, families, and households.

Although survey data are still the backbone of social science research, they are better integrated than before. Recent work has developed algorithms and matching software to link person records across datasets or link surveys to other public administrative data sources. For example, Social Security Administration, partnered with the Census Bureau, has linked Social Security benefit and earnings data as well as IRS income files with two of the major U.S. surveys, the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP) (McNabb et al. 2009). The Panel Study of Income Dynamics (PSID), the longest-running longitudinal household survey in the United States, has linked the survey data with external mortality and health data from the National Center for Health Statistics and Centers for Medicare and Medicaid Services, housing subsidy data from the Department of Housing and Urban Development, and school performance data from the National Center for Education Statistics (McGonagle et al. 2012). This project is

part of a broader effort to link U.S. Census to several other major aging and life course surveys, including the Health and Retirement Study (HRS), the Wisconsin Longitudinal Study (WLS), the National Social Life, Health, and Aging Project (NSHAP), and the National Health and Aging Trends Study (NHATS) (Warren et al. 2020). These data linkage infrastructures not only have important potential to improve cutting-edge social science research, but also facilitate policy designs and evaluations.

Other Big Data

As we mentioned earlier, administrative data are one form of big data that have changed the landscape of social science and policy research. With the internet and increased computational power, many other types of big data produced as by-products of commercial or social transactions also enrich data sources that can be used for academic research. Examples of well-known big data include Google search results, social media tweets, traffic camera records, mobile phone location tracking, internet site browsing history, and many other digital data in the forms of texts, photos, audios, and videos. Organic big data are not the focus of this essay but require mention for many reasons. Such data have and will continue to transform business and have the potential to provide valuable research insights.

Many benefits of administrative data also apply to organic big data: large samples with high-quality information (when properly used). Data are often produced at high frequency and in real-time, providing sample sizes undreamed of in traditional social science research but requiring new computational and statistical techniques. Data are often less structured (or with more complex structure) than traditional cross-sectional or panel data, thus requiring new approaches to data storage and retrieval. Although such organic

big data can be hardly linked with administrative or survey data at the individual level, they can provide aggregate information about certain neighborhoods, areas, and hard-to-reach geographic regions. For example, Alexander, Polimis, and Zagheni (2020) show how to combine social media and survey data to estimate migration flows in the United States. Basellini et al. (2020) illustrate the linking of COVID-19 mortality data from national statistics with Google mobility data. Although linking organic big data and administrative data abounds with new benefits and challenges, creative use of these data will potentially break new ground on old questions and broaden the horizon of scientific inquiry.

Conclusion

New sources of administrative data show promise, but they complement rather than supplant traditional surveys. The data revolution brings new challenges to social and policy research. Future studies will need to focus on better integration of different data sources through developing joint protocols and methodologies; promoting the accumulation of knowledge; bringing together tools from mathematics, statistics, and computer science; making use of theory, algorithms, mechanisms, and practices; and addressing social, legal, and ethical considerations. Administrative microdata are vital for understating underlying trends in economic development, marriage and family transitions, urbanization, internal and international migration, and aging and population health. Similar data-linking projects for public use are emerging in many other countries and provide unprecedented opportunities to expand the scope of social science research and develop policies to solve pressing social problems.

References

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2020. "Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States." arXiv preprint arXiv:2003.02895.

Auten, Gerald and David Splinter. 2019. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." Draft subject to change. http://davidsplinter.com/AutenSplinter-Tax_Data_and_Inequality.pdf.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey. 2019. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." No. w24019. National Bureau of Economic Research.

Basellini, Ugofilippo, Diego Alburez-Gutierrez, Emanuele Del Fava, Daniela Perrotta, Marco Bonetti, Carlo Giovanni Camarda, and Emilio Zagheni. 2020. "Linking Excess Mortality to Google Mobility Data during the COVID-19 Pandemic in England and Wales." Working Papers, The French Institute for Demographic Studies. https://hal.archives-ouvertes.fr/hal-02899654

Card, David, Raj Chetty, Martin S. Feldstein, and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." SSRN Scholarly Paper ID 1888586, Social Science Research Network, Rochester, NY.

Connelly, Roxanne, 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* 59:1–12.

Ferrie, Joseph, Catherine Massey, and Jonathan Rothbaum. 2016. "Do Grandparents and Great-Grandparents Matter? Multigenerational Mobility in the US, 1910–2013." Technical report, National Bureau of Economic Research.

Goerge, Robert M. and Bong Joo Lee. 2002. "Matching and Cleaning Administrative Data." In *Studies of Welfare Populations: Data Collection and Research Issues*, edited by Michele Ver Ploeg, Robert A. Moffitt, and Constance F. Citro, pp. 197–219. Washington, DC: National Academy Press.

Groen, Jeffrey A. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28:173–198. https://www.scb.se/dokumentation/statistiska-metoder/JOS-archive/.

Groves, Robert M. 2004. *Survey Errors and Survey Costs*. Hoboken, New Jersey: Wiley-Interscience.

Groves, Robert M. 2011. "Designed Data and Organic Data." Technical report, U.S. Census Bureau. Library Catalog: www.census.gov Section: Government.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology (Second Edition)*, volume 561. John Wiley & Sons.

Grusky, David B, Michael Hout, Timothy M Smeeding, and C Matthew Snipp. 2019. "The American Opportunity Study: A New Infrastructure for Monitoring Outcomes, Evaluating Policy, and Advancing Basic Science." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5:20–39.

Guvenen, Fatih and Greg Kaplan. 2017. "Top Income Inequality in the 21st Century: Some Cautionary Notes." Working Paper 23321, National Bureau of Economic Research.

Massey, Catherine G, Katie R Genadek, J Trent Alexander, Todd K Gardner, and Amy O'Hara. 2018. "Linking the 1940 US Census with Modern Data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51:246–257.

McGonagle, Katherine A, Robert F Schoeni, Narayan Sastry, and Vicki A. Freedman. 2012. "The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research." *Longitudinal and Life Course Studies* 3:268–284.

McNabb, Jennifer, David Timmons, Jae Song, and Carolyn Puckett. 2009. "Uses of Administrative Data at the Social Security Administration." *Social Security Bulletin* 69 (1):75–84.

Medalia, Carla, Bruce D. Meyer, Amy B. O'Hara, and Derek Wu. 2019. "Linking Survey and Administrative Data to Measure Income, Inequality, and Mobility." *International Journal of Population Data Science* 4 (1): 1–8.

Meyer, Bruce D., and Nikolas Mittag. 2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net." *American Economic Journal: Applied Economics* 11 (2): 176–204.

OMB. 2020. "Request for Comment on Considerations for Additional Measures of Poverty." Regulations.Gov. April 14, 2020. https://www.regulations.gov/docket?D=OMB-2019-0007.

Piketty, Thomas and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *The Quarterly Journal of Economics* 118:1–41.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51: 287–297.

Ruggles, Steven, Robert McCaa, Matthew Sobek, and Lara Cleveland. 2015. "The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata." *Journal of Demographic Economics* 81:203–216.

Slemrod, Joel. 2016. "Caveats to the Research Use of Tax-Return Administrative Data." *National Tax Journal* 69: 1003–1020. Publisher: National Tax Association.

United Nations. 2007. *Register-Based Statistics in the Nordic Countries—Review of Best Practices with Focus on Population and Social Statistics*. Methodological Guidelines. UN. http://digitallibrary.un.org/record/609979.

Warren, John Robert, Fabian T Pfeffer, Jonas Helgertz, and Dafeng Xu. 2020. "Linking 1940 US Census Data to the Panel Study of Income Dynamics: Technical Documentation." Technical report, University of Michigan PSID Technical Series Paper 20-02.