Expanding Access to Administrative Data for Research in the United States*

David Card, UC Berkeley
Raj Chetty, Harvard University
Martin Feldstein, Harvard University
Emmanuel Saez, UC Berkeley

Abstract

We argue that the development and expansion of direct, secure access to administrative micro-data should be a top priority for the NSF. Administrative data offer much larger sample sizes and have far fewer problems with attrition, non-response, and measurement error than traditional survey data sources. Administrative data are therefore critical for cutting-edge empirical research, and particularly for credible public policy evaluation. Although a number of agencies have successful programs to provide access to administrative data – most notably the Centers for Medicare and Medicaid Services – the United States generally lags far behind other countries in making data available to researchers. We discuss the value of administrative data using examples from recent research in the United States and abroad. We then outline a plan to develop incentives for agencies to broaden data access for scientific research based on competition, transparency, and rewards for producing socially valuable scientific output.

A Wealth of Administrative Data

Governments create comprehensive micro-economic files to aid in the administration of their tax and benefit programs. The Social Security Administration (SSA), for example, records annual data on earnings and retirement and disability benefit payments for virtually the entire US population. State agencies collect quarterly earnings reports from firms on behalf of the Department of Labor for nearly all paid workers in the private sector. The Internal Revenue Service and the various state income tax administrations compile income data for all individuals and businesses. The Medicare and Medicaid programs record information on the health care services received by their beneficiaries. School districts record detailed information on academic outcomes, classes and teachers for all public school students. Counties record every real estate transaction. A rich archive of information covering most aspects of socio-economic behavior from birth to death, including education, earnings, income, workplace and living place, family composition, health and retirement, is recorded in administrative data.

With the advent of modern computer systems, all these administrative data are stored in electronic files that can be used for statistical analysis. Indeed, government agencies are required to produce statistical reports that inform the public about their activities, and hence have already established statistical offices and set up the necessary files to produce such information. Each year, information for virtually every person in the country is used multiple times in the production of official US statistics, while maintaining the strictest standards of privacy.

Eroding US Leadership

Traditionally, empirical research in social sciences has relied on survey data sources such as the decennial Census, the Current Population Survey (CPS), the Panel Study of Income Dynamics (PSID), and the Survey of Income and Program Participation (SIPP). In the post-war period the US led the way in the development of modern survey methods, and not coincidentally, in the development of statistical

* This paper was written for the National Science Foundation 10-069 call for white papers on "Future Research in the Social, Behavioral & Economic Sciences." We thank Nick Greenia, Alan Krueger, Julia Lane, Nancy Lutz, Alex Mas, Dan Newlon, Matthew Shapiro, and Joel Slemrod for helpful comments and suggestions. A shorter and less documented version of this paper is posted on the NSF website.

techniques for analyzing these data. The combination of data and methods established the nation's dominant position in the conduct of empirical social science research. During the second half of the 20th century, the fields of political science, sociology, and economics were all revolutionized by US researchers using US-based survey data sources.

Unfortunately, that dominant position is now at risk as the research frontier moves to the use of administrative data. Administrative data are highly preferable to survey data along three key dimensions. First, since full population files are generally available, administrative records offer much larger sample sizes. The full population earnings data from SSA or tax records is about 2000 times larger than the CPS. Larger sample sizes can be harnessed to generate more compelling research designs and to study important but relatively rare events – like a plant downsizing that affects some workers but not others, or a severe local weather event. Second, administrative files have an inherent longitudinal structure that enables researchers to follow individuals over time and address many critical policy questions, such as the long term effects of job loss (von Wachter, Song, and Manchester, 2009), or the degree of earnings mobility over the life cycle (Kopczuk, Saez, and Song, 2010). Third, administrative data provide much higher quality information than is typically available for survey sources, which suffer from high and rising rates of non-response, attrition, and under-reporting. In the recent CPS Annual Social and Economic Supplements, for example, 31% of earnings, 34% of Social Security and retirement income, and 30% of public assistance amounts have to be imputed due to non-response (Nelson, 2006). The rate of attrition from the SIPP, which is designed to measure changes in income and program participation, exceeds 30% (Slud and Bailey, 2006). Even with imputations, only about 60% of the welfare income paid to individuals is reported in the CPS (Meyer and Sullivan, 2006). Under-reporting in SSA earnings data or income tax data is clearly much lower.

Because of confidentiality and security concerns, administrative data cannot be made publicly available. However, numerous examples -- from the Centers for Medicare and Medicaid Services (CMS), from other countries, and from a variety of pilot efforts at federal, state, and local government agencies -- show that it is possible to provide secure access to de-identified administrative data (i.e., data that have been stripped of individual identifiers such as names, addresses, and social security numbers) to researchers. To the best of our knowledge, research access to de-identified data has never resulted in the improper disclosure of confidential information. The record shows that access can be achieved in a way that maintains the strictest standards of privacy while still allowing researchers direct access to individual records.

A leading example of the research impact of routine access to administrative micro-data is CMS. Many hundreds of medical studies each year use the agency's Research Data Assistance Center (ResDAC) to develop requests for micro data files (including data protection plans), which are then reviewed by CMS. Routine access to Medicare and Medicaid files has enabled US healthcare researchers to maintain their global leadership position in the field and have yielded many important public benefits.

Outside the US, many countries have developed systems to allow access to administrative data for research purposes. In Denmark for example, Statistics Denmark gives prepares de-identified data by combining information from administrative databases for approved research projects. The data extracts can then be accessed by researchers remotely (from any computer, including the researcher's office desktop) through a secure server. Researchers apply for data access through accredited "centers" at major universities, and access is provided through an open competition process based on scientific merit.

The availability of detailed administrative data abroad has led to a shift in the cutting edge of empirical research in many important areas of social science away from the United States and toward the countries with better data access. Because the US retains worldwide leadership in the quality of its academic researchers, US-based researchers are often involved in research using administrative data from other

countries (e.g., Card, Chetty, and Weber, 2007 on the effects of unemployment insurance; Chetty et al. 2011 on labor supply and taxes). However, this situation is less than ideal for at least two reasons.

First and most important, many questions of central importance for US policy making cannot be tackled using evidence from other countries. The decentralized US labor market is quite different from the European labor market (where workers are often covered by collective contracts). Other US institutions are also fundamentally different: most European countries still allow mandatory retirement; many have individual rather than family-based taxation; and most provide a very different combination of income support and welfare programs. Hence, evidence from job creation programs in Europe cannot be easily applied to the United States. Access to existing national data, however, could easily be used to evaluate the effects of various US government policies, such as stimulus spending, on job creation and overall personal income. US public policy would be far better served having top researchers focusing on US policies issues using US data, but this requires the development of administrative data access.

Second, in the long-run, the development of administrative data access abroad will foster the development of empirical and econometric research programs in those countries, in the same way that the development of US survey data was accompanied by great scientific progress in empirical methods in social sciences in the United States in the 20th century. The United States would be much better positioned in this scientific race if access to US administrative data was the norm rather than the exception in limited pilot programs.

Regaining US Leadership

Over the years, the United States has developed a number of initiatives to provide access to administrative data access for research. As discussed above, data access is relatively good for health data, especially for people covered by Medicare. Similarly, hospital records and birth and death records can be accessed in many states, and have been used to evaluate many critical policy issues (including environmental policy issues). The situation with respect to K-12 education is also relatively good. Some large school districts have established co-operative programs with researchers (e.g., Chicago); some states have made anonymized student record data readily available (e.g., North Carolina), and a few states have begun to allow researchers to link school records and labor market data from UI records (e.g., Florida).

Access to data on income and earnings is not as satisfactory, although some valuable initiatives exist. In principle, unemployment insurance records for many states can be accessed through the LEHD program at the Census Bureau – although this is onsite at a Census RDC. In recent years, SSA earnings data have been accessed by researchers through IPAs (internships in the SSA Washington DC offices) or co-authorship with SSA researchers. The Statistics of Income division of the US Treasury has also launched a promising tax data access program for statistical research purposes. In all these cases, however, the lack of sufficient resources and the "bandwidth" to accommodate many simultaneous research projects, as well as cumbersome application and data access procedures, severely limit the research potential.

Based on experiences from other countries and these pilot initiatives, we believe that five conditions must be satisfied to make a data access program sustainable and efficient:

(a) fair and open competition for data access based on scientific merit
(b) sufficient bandwidth to accommodate a large number of projects simultaneously
(c) inclusion of younger scholars and graduate students in the research teams that can access the data
(d) direct access to de-identified micro data through local statistical offices or, more preferably, secure remote connections[1]
(e) systematic electronic monitoring to allow immediate disclosure of statistical results and prevent any disclosure of individual records

[1] Many federal government employees are now able to access confidential data at home via secure "flexiplace" systems, showing that it is feasible to implement such remote access systems in the U.S.

We emphasize that **direct access to micro-data** is critical for success. Alternatives such as access to synthetic data or submission of computer programs to agency employees will **not** address the key problem of restoring US leadership with cutting-edge policy-relevant research.

Synthetic data is simulated micro data that is constructed to mimic some features of the actual data. This approach is much less attractive than providing direct access to the full administrative data set because in practice it is virtually impossible for the researchers to fully specify the contents of the ideal synthetic dataset in advance. Moreover, many important policy questions require information on the entire distribution of outcomes of interest, and the methods that have been recently developed to address these questions (such as quantile treatment effect methods) require access to the full sample of actual data. Synthetic data also make it difficult or impossible to study subpopulations, unless distinct synthetic data sets are specifically created for this purpose. This approach is a very poor substitute for authorized secure access to actual administrative record micro-data files, and will require a large infrastructure of intermediaries whose job is to construct synthetic data sets on a project-by-project basis, effectively adding noise to the existing data, which researchers will then try to remove using advanced statistical methods which may or may not work. Experience shows that researchers are almost always best served when starting from the raw data. Any processed data in general requires time consuming reverse engineering to undo or control for the processing noise.

The option of sending computer programs, while providing some data access, is also substantially inferior to direct data access because it does not allow for the inductive phase of data analysis that is critical for many empirical projects.[2] Empirical researchers learn from examining and analyzing data records directly, and from simple summary statistics for subsets of data records, and often discover anomalies or problems with the data (or their hypotheses) that would be missed by simply examining regression outputs. Without direct access to the micro data this critical step in the scientific process is largely missing.

The Value of Competition

In principle, having a centralized agency being able to obtain administrative data from all government branches and then maintain it and supply de-identified data to approved research projects, as in the Danish case, is an attractive model. However, in the US, this model is less attractive for four reasons. First, relative to other countries, the US government is far more decentralized, with multiple agencies at three different levels covered by statutory limits on inter-agency data sharing. In some cases the creation of a central warehouse would require legislative action to amend these statutes.[3] Second, and closely related, different agencies are covered by different privacy laws and have to satisfy different requirements to meet the standards for disclosure of private data for research purposes. Third, there is a long tradition of distrust of centralized government in the US, and in particular of monopoly control by a single government agency. Any successful data access program must acknowledge the salience and value of this tradition. Finally, from the perspective of both privacy and efficiency, it would seem reasonable to leverage the existing statistical offices of US administrative agencies for both their expertise and also as a base for access to such confidential data.

[2] For example, the top wealth share analysis of Kopczuk and Saez (2004) was done by sending computer programs to Barry Johnson at the Statistics of Income division at the US Treasury. This was feasible in this case only because a relatively simple set of tabulations were required for the project, and the data were already in excellent shape for research.

[3] As of today, even the Census Bureau which is the central US statistical agency, has only very limited access to tax data, even for the production of official US statistics, although efforts at improving data sharing have taken place for decades.

We therefore believe that it is preferable to leverage the multiple agency setting and the principle of inter-agency competition by allowing and encouraging different agencies to provide their own data access systems. This could be achieved by rewarding agencies for performance. Performance in scientific production is easily measurable via metrics such as peer-reviewed publications. Rewards to agencies could take the form of resources provided by the major research funders (NSF and NIH) that would help agencies strengthen their statistical offices and develop partnerships with researchers. Currently, the main hurdle in the development of research partnerships between agencies and external researchers is the lack of internal incentives and the lack of dedicated agency resources. A well designed system would encourage agencies to improve their statistical capabilities and data access, subject to agency-specific rules that ensure the strictest standards of privacy. This model – which closely parallels the model of the Centers for Medicare and Medicaid Services -- is much more robust than the centralized agency model, and would unleash the forces of innovations as agencies compete for the best research projects. This model can also be extended to private institutions that gather data valuable for research (such as utilities for the analysis of energy and resource conservation for example) to create incentives for research partnerships.[4] Both government agencies and private institutions already have multiple business contracts for data work where outside contractors access the data for a specific business purpose. Scientific research should follow the same model where NSF or NIH funds researchers to carry out scientific projects with the data.

The Value of Cooperation

Experience from abroad and from the United States shows that there is tremendous value in carrying research by merging data, for example educational data and earnings data. A centralized agency, as in Denmark, naturally allows such merging. However, starting from the decentralized landscape we have described, it should be possible to encourage partnerships between two government statistical agencies (or between a statistical agency and an external partner such a non-profit or business) to accommodate research requiring merged data. Such cooperation will naturally arise if all parties can share the benefits of the scientific output. Precedents for this kind of cooperation exist even in the US. Recently, the Florida Department of Education has teamed up with the state UI agency to allow linking of student education records to subsequent earnings outcomes.[5]

Another important example is the long-term analysis of randomized field experiments. Field experiments are a powerful method for scientific evaluation of alternative policy choices, and the US was an early leader in the use of field experiments to evaluate negative income tax policies in the 1960s. During the late 1980s and early 1990s, many states tested innovations to their welfare programs under the so-called "Waiver" program, which required randomized evaluation. In most cases these evaluations relied on administrative earnings data (see Harvey, Camasso, and Jagannathan, 2000). Recently, Chetty et al. (2010) used tax data to evaluate the effects of Project STAR – one of the largest scale experiment on K-3 early childhood education which was conducted in the 1980s – on long term student outcomes, including college attendance and earnings. In that case, it was possible to merge information from the STAR experiment to tax data through the Statistics of Income external researcher program at the US Treasury.

Recently, there has been renewed interest in the use of field experiments in many areas of economics. The ready availability of access to US administrative data would allow researchers to overcome difficulties of

[4] There are already examples of direct partnerships between researchers and private businesses. For examples, a series of influential studies on retirement savings behavior has been carried out using data from private pension funds management companies (Madrian and Shea, 2001).

[5] Even in cases where data cannot be shared with other agencies due to statutory limitations, it is nevertheless possible to bring the least protected data into the agency maintaining the most protected data. That is why developing data access to the most protected data, such as tax data, is so critical, as any other form of data can later be merged into tax data.

tracking, non-response, and under-reporting in conventional survey-based measures, and allow the analysis of long-term outcomes, hence substantially expanding the scientific value of these experiments at relatively low cost. The development of access to administrative data would put the United States back in the forefront of this new wave of scientific research.

References

Card, David, Raj Chetty, and Andrea Weber (2007). "Cash-on-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market." *Quarterly Journal of Economics* 122(4): 1511-1560.

Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri (2011). "Adjustment Costs, Firm Responses, and Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics*, forthcoming.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. (2010). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." NBER Working Paper 16381.

Harvey, Carol, Michael J. Camasso, and Radha Jagannathan (2000). "Evaluating Welfare Reform Waivers under Section 1115." *Journal of Economic Perspectives* 14(4): 165-188.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song (2010). "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937." *Quarterly Journal of Economics*, 125(1): 2010, 91-128.

Kopczuk, Wojciech and Emmanuel Saez (2004). "Top Wealth Shares in the United States, 1916-2000: Evidence from Estate Tax Returns." *National Tax Journal* 57(2), Part 2: 445-487.

Madrian, Brigitte C. and Dennis F. Shea (2001)."The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116(4): 1149-1187.

Meyer, Bruce and Francis X. Sullivan (2006). "Consumption, Income, and Material Well-Being After Welfare Reform." Unpublished Manuscript, University of Notre Dame, January.

Nelson, Charles (2006). "What Do We Know About Differences Between CPS and ACS Income and Poverty Estimates? US Bureau of the Census Housing and Household Economic Statistics Division. Unpublished Memorandum, August 21.

Slud, Eric V. and Leroy Bailey (2006). Estimation of Attrition Biases in SIPP." Paper presented at the ASA Section on Survey Research Methods.

Von Wachter, Till, Jae Song and Joyce Manchester (2009). "Long-Term Earnings Losses due to Mass-Layoffs During the 1982 Recession: An Analysis Using Longitudinal Administrative Data from 1974 to 2004." Unpublished Working Paper, April.