January 5, 2017
Chicago, Illinois

Testimony before the Commission on Evidence-Based Policymaking

by

V. Joseph Hotz
Arts & Sciences Professor of Economics
Duke University
Durham, NC 27708
hotz@econ.duke.edu

Commissioners Abraham, Haskins, Meyer and Troske, it is a pleasure to testify before you this afternoon.

The testimony I provide today is related to comments submitted on behalf of the Population Association of America (PAA) and the Association of Population Centers (APC) on November 14, 2016. That said, I am not here today to speak on behalf of either organization. All of the views and opinions expressed below are my own and ones I support.

In my comments today I will focus on two related issues: the data infrastructure to support evidence-based policy analysis and the importance of insuring access to these data for qualified researchers, both those within and outside of government. In the interest of time, I will forego discussing some of the other issues the Commission will address, including appropriate designs for impact and implementation evaluations of public programs and the important issue of protecting the privacy and confidentiality of data used in such evaluations that are collected from U.S. citizens, firms and other institutions and entities.

I will organize my remarks into four broad topics related to data and data access.

A. *There are important benefits to the use of not only governmental administrative data, but increasingly to the use of non-governmental sources for conducting policy-relevant research, especially when these data are linked together and/or with survey data.*

I don't need to tell the members of this Commission of the essential role that administrative records have played in a variety of areas of important policy-relevant research. For example, such records have been used to study participation in and impacts of *social programs* (e.g., welfare programs, manpower training, food stamps, the EITC, etc.) on various *outcomes*. Often, the outcomes of interest are measured with administrative data – such as wage earnings (obtained from linked unemployment insurance wage records), health conditions (obtained from linked Medicaid records) or fertility-related outcomes (using linked birth certificate records) – *linked to caseload or participant records*. The availability of administrative records from federal, state or local sources provide a cost-effective way of supporting evaluations of these programs, regardless of whether the evaluations make use of randomized or non-experimental designs for assignment of treatment.

But, increasingly, non-governmental sources of administrative data are playing an important role in monitoring policy-relevant issues and/or evaluating particular policies or "treatments." Here I reference two examples.

First, biomedical research, including research that is relevant to policies affecting health-related behaviors, such as smoking bans or regulation of the nutritional content of foods, increasingly make use of electronic health records (EHRs) from public and private health care systems to measure the health effects of variation in such policies.

Second, administrative records from private firms that construct credit scores for use by financial institutions have been used by researchers, including the research division of the New York Federal Reserve Bank, to monitor and conduct policy-relevant research on student loan debt in the U.S.

These examples illustrate the benefits that non-governmental administrative data can have in conducting policy-relevant research, including their cost-effectiveness and potential for more accurately measuring a variety of phenomena. And, people like Commissioner Robert Groves know, they are the tip of the iceberg, given the increasing availability of "big data" that can be scraped from the Internet. *I urge the Commission to note the importance of non-governmental admin records in assessments of sources of data for evidence-based policymaking*.

B. ***There are important legal and other constraints that limit the use of governmental administrative records and the ability of researchers to link records across different sources of these data.***

In particular, different sources of governmental administrative data are subject to varying and divergent laws and regulations that can inhibit access to them. For example, administrative records from social programs administered at the state or local level (e.g., TANF programs) are often subject to laws and regulations that make it hard for one agency to share their records with another agency. And, as noted in the NRC report on the Reengineering of the Survey of Income and Program Participation (SIPP), existing state laws that cover the privacy and access of administrative records from TANF, Medicaid, unemployment insurance, and the workers' compensation programs make it very difficult, if not impossible, for these programs to link their data with the Census Bureau (or other) surveys like the SIPP.[1]

Historically, this issue has complicated the conduct of biomedical research that makes use of electronic (or non-electronic) health care records of individuals as institutional review boards (IRBs) have required studies to obtain informed consent from subjects in these studies for any follow-up use of subjects' EHRs and/or updating of these records. Recently proposed revisions to the Common Rule[2] will reduce and/or eliminate this re-consenting requirement for certain types of studies and types of administrative records so long as subjects are provided with a clear statement regarding potential future use of administrative records as part of their initial consent process.

---

[1] Constance Citro and John Karl Scholz, eds., *Reengineering the Survey of Income and Program Participation*, National Research Council, 2009. [NOTE: I served as a member of the NRC expert panel that developed this report.]

[2] HHS–OPHS–2015–008, Federal Policy for the Protection of Human Subjects, *Federal Register*, 80(173), Sept 8, 2015, 53933-54058.

Many population scientists welcome this change and suggest it may represent a model for the Commission to examine as it considers how to facilitate access to records like EHRs while still providing participants with the opportunity to make informed decisions about research access to their records.

I urge the Commission to investigate the various laws and regulations governing access to administrative records for research purposes. In particular, I encourage the Commission to look closely at the laws affecting access to state and local government data and policies restricting record linkage across various federal agencies as part of your deliberations.

C. *To facilitate the conduct of evidence-based, policy-relevant research, I encourage the Commission to examine and seek to improve the access to governmental administrative records by qualified researchers inside and outside the government.*

I understand and appreciate that there are important confidentiality and security concerns that necessarily limit access of researchers to various types of government-based administrative records and/or other restricted-use data sources. Furthermore, I appreciate why restrictions on access by non-governmental researchers may need to be different, and possibly more restrictive, than that applied to researchers employed by authorized governmental agencies. But, at times, these restrictions have made access to such data very difficult for academic and non-governmental researchers, researchers at other levels of government and those from the non-profit and profit-making sectors.

Over the last 20 years, U.S. statistical agencies, initially led by the U.S. Census Bureau, have made great strides in improving access to restricted-use versions of federal data sources through the Federal Statistical Research Data Centers (RDCs) program. This program now allows access to data products from 12 different federal statistical agencies for qualified governmental and non-governmental researchers in 20 different centers around the country. While some the research covered by the data agreements conducted in these centers is often not directly related to policymaking, much of it is.

A similar effort for providing access to data from the Internal Revenue Service (IRS) under the Joint Statistical Program of the Statistics of Income (SOI) Division of the IRS has enabled qualified researchers to submit proposals for access to IRS data and to link it to various data for research purposes. This program has facilitated a number of highly visible and widely cited lines of research by Professors Raj Chetty (Stanford) and Emmanuel Saez (UC Berkeley) and their collaborators. For example, Chetty and co-authors have analyzed the association between income and the life expectancy of individuals in the U.S. since 2000 by linking IRS tax records on income with Social Security Administration death records.[3] The findings of this research, especially the finding of differences by geography in the association between mortality and income, raises important questions about the reasons of these disparities and how to alleviate them. Such research could not have been conducted without this data access program.

The work by Chetty et al. is part of a large body of research that shows that geography (e.g. neighborhoods) affects the social and economic well-being and health of individuals and families. Better understanding the "mechanisms" behind these neighborhood effects is important for public

[3] Chetty, R. et al. (2016). "The Association between Income and Life Expectancy in the United States, 2001-2014, *Journal of the American Medical Association*, 315(15): 1750-1766.

policy. But, state and local policymakers, researchers, and program officials often lack the data needed to measure differences in community environments, to isolate how neighborhood characteristics shape micro-level outcomes, or to test the efficacy of neighborhood-level interventions. Most survey data files lack such key contextual information, while most administrative data lack key demographic, socioeconomic, behavioral, and outcome information. While individual-level record linkage of survey and administrative data could provide such critical data for state and local-level evidence-based policymaking, most state and local researchers/program evaluators lack the resources to submit proposals and conduct these types of linkages and research within a RDC.

I hope that the Commission will seek to encourage expanding access of data and records from federal, state and local sources to qualified governmental and non-governmental researchers, including state and local government researchers. This should include expanding access and easing the requirements for such access to the RDC and/or IRS's Joint Statistical Programs or similar programs. I also encourage the Commission to examine other options for providing researchers with greater access to sensitive data outside of RDCs, including the release of *de-identified and/or synthetic* versions of fine-geography data and linked versions administrative and survey data.

I also encourage the Commission to consider recommending revisions to statutory or administrative laws that would allow federal statistical agencies to share data with researchers conducting evidence-based research. For example, the Census Bureau's authorizing regulation, Title 13, does not explicitly recognize the use of sensitive data for conducting scientific research, be it policy-relevant or not, as a "benefit to the Bureau." Rather, Title 13 only supports data access to improve the quality of Census data products. A more explicit acknowledgment that qualified research using these data are beneficial would enable the Census Bureau to approve studies primarily designed to replicate existing studies and/or determine the robustness of findings from previous research. *Such changes will encourage the replicability of findings from policy-relevant research, something that should be a core principle for evidence-based policymaking and not just a principle of scientific inquiry.*

D. *I encourage greater attention be given to the population representativeness of the policy-relevant research produced using data from administrative records and/or surveys.*

Many studies use administrative records to "evaluate" the impact of a particular policy or program. As I argued earlier, administrative records provide a potentially cost-effective way of conducting such evaluations. But the benefits of using administrative records in evaluative research does not mitigate the importance of assessing the sampling properties of these data sources and their consequences for the generalizability of the findings they produce.

Consider the following example from bio-medical research – which I include under the heading of *policy-relevant health research* – regarding the design of the Precision Medicine Initiative (PMI). One of the key components of the PMI's initial plan is to assemble a million-person sample of individuals who would provide access to their Electronic Health Records (EHRs) as a condition of the study. Access to EHRs on this large sample would provide data to study a wide range of health conditions, including conditions that are relatively rare and only affect population subgroups. One of the study's recruitment strategies was to use social media and other methods to attract participants who would grant access to their EHRs and undergo one or more physical examinations.

While the goals of the PMI are important and have the potential to provide evidence-based

assessments of health conditions relevant for U.S. health policy, population scientists and other social scientists are concerned about lack of attention to the properties of what amounts to a "volunteer" sample of people with EHRs, even if the sample includes data on one million participants. In public comments, Population Association of America and the Association of Population Centers have raised these concerns and strongly suggested that the NIH leadership consider using existing population-based health studies to form at least part of the PMI cohort to assess the *population-representativeness* of the recruitment strategy based on volunteers.

In developing both policies and best practices for policy-relevant research, I encourage the Commission to advocate for the designs of data collection that explicitly account for the sampling properties and population-representativeness of these data and of the findings they produce.

Lastly, *I encourage the Commission to ensure that population-representative data sources collected by the Federal government continue to be viewed as an important source of data for policy-relevant research, both as a way to monitor behaviors and phenomena relevant to public policy.* For example, data sources like the Current Population Survey (CPS), the American Community Survey (ACS) and the Survey of Income and Program Participation (SIPP) all play roles in the monitoring and implementation of a variety of public policies in the U.S. The CPS is the population-representative data that enables the BLS to construct estimates of unemployment and labor force participation rates of the U.S. population on a monthly basis. The ACS provides data on poverty rates at the lowest levels of geography, such as school districts and communities, which are used to allocate funding for programs such as the USDA's National School Lunch Program and State Children's Health Insurance Program (SCHIP). The SIPP has facilitated a broad range of research on the distribution of income and participation in a range of social programs using a survey that is designed to be population representative for most states in the U.S. These surveys, and others, are important components of the U.S. data infrastructure and are needed to support evidence-based policymaking.

I thank you for the opportunity to present these views to the Commission and am happy to answer any additional questions you may have.