V. Joseph Hotz

Professor of Economics, Duke University

Professor Hotz specializes in the subjects of applied econometrics, labor economics, economic demography, and economics of the family. His studies have investigated the impacts of social programs, such as welfare-to-work training; the relationship between childbearing patterns and labor force participation of U.S. women; the effects of teenage pregnancy; the child care market; the Earned Income Tax Credit; and other such subjects.

He began conducting his studies in 1977, and has since published his work extensively in books and leading academic journals. Many of his projects have been funded by grants awarded by the National Institute of Health and the National Science Foundation. He is currently completing a project with Duncan Thomas on, “Preference and Economic Decision-Making” under a grant from the National Institute of Child Health and Human Development. His recent works also include, “Tax Policy and Low-Wage Labor Markets: New Work on Employment, Effectiveness and Administration” with John Karl Scholz and Charles Mullin; and “Designing New Models to Explain Family Change and Variation” with S. Philip Morgan.

Along with his duties as an independent researcher, Professor Hotz has also held positions as a research associate of the National Bureau of Economic Research, the National Poverty Center, the Institute for the Study of Labor, and the Institute for Research on Poverty. He is presently a member of the Committee on National Statistics for the National Academy of Sciences’ Research Council.

December 20, 2016

Statement and Recommendations Concerning
Commission on Evidence-Based Policymaking

by

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

The following comments are related to comments submitted on behalf of the Population Association of America (PAA) and the Association of Population Centers (APC) on November 14, 2016. Nonetheless, all of the views and opinions expressed below are my own and ones I support.

Let me begin by applauding passage of the Evidence-Based Policymaking Commission Act of 2016, which created the bipartisan Commission on Evidence-Based Policymaking and charged this Commission with:

- determining how to integrate administrative and survey data and to make those data available to facilitate research, evaluation, analysis, and continuous improvement while protecting privacy and confidentiality;

- recommending how data infrastructure, database security, and statistical protocols should be modified to best fulfill the integration and increased availability of data as described above;

- recommending how best to incorporate rigorous evaluation into program design; and

- considering whether a federal clearinghouse should be created for government survey and administrative data.

In what follows, I focus my comments 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 within and outside of government.

A. ***There are important benefits to the use of administrative data, especially when linked, for conducting policy-relevant research.*** Administrative records have been used in a variety of research areas and provide an essential source of data for conducting 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 earned income tax credit, etc.) on various outcomes. Often these outcomes are measured with linked administrative data, such as wage earnings (from linked unemployment insurance wage records), health conditions (from linked Medicaid records) or fertility (from linked birth certificate 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 made use

of randomized designs for allocation of program participants to different "treatments," or other studies that have made use of non-experimental designs.

But social program evaluation is not the only place where administrative records can and will be the primary source of data to monitor particular programs and/or evaluate particular policies or "treatments." Furthermore, they do not only use government records. 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, uses increasingly 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. In both of these areas, administrative records support important policy-relevant research in a way that is both cost-effective and potentially more accurate than data collected via other means, e.g. surveys.

B. ***At the same time, there are important legal and other constraints that limit the use of administrative records and the ability of researchers to link records across different sources of these records.*** In particular, different sources of administrative data are subject to varying and divergent laws and regulations that can inhibit their use. 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 share 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. 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.

More generally, I strongly urge the Commission to investigate the various laws and regulations

---

[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.

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.

C. ***To facilitate the conduct of evidence-based, policy-relevant research, I encourage the Commission to examine and improve access to administrative records to qualified researchers outside and inside 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 restricted-use data sources. Furthermore, I appreciate why restrictions on the access of non-governmental researchers may need to be different, and possibly more restrictive, than that applied to researchers employed by authorized government agencies. But, at times, these restrictions have made access to such data very difficult for academic and non-governmental researchers.

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 approved for use of 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). For example, Chetty and co-authors 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 area in the associations of longevity by income, raises important questions about the sources of these disparities and how to alleviate these differences. Such research could not have been conducted without this program.

A large body of research shows that geography (e.g. neighborhoods) affects the social and economic well-being and health of individuals and families. 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. The

[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.

Commission should also encourage statistical agencies and other researchers to create spatially-linked administrative and survey data that could be provided to state and local researchers/program evaluators outside of RDCs to increase evidence-based policymaking at the state and local level.

I call on the Commission to encourage expanding access of data and records from federal, state and local sources to qualified non-governmental and governmental researchers. This expansion should include state and local government researchers, whose access to data can provide support for accurately assessing needs, creating programs to address those needs, and delivering services in more cost-effective ways. While such efforts may include expanding the RDC and/or IRS's Joint Statistical Programs or similar programs, they should also include expanding access to spatially-linked administrative and survey data that could be provided outside of RDCs. Efforts also may include providing more funding for merit-based grants to undertake these projects, especially in light of the limited resources available to researchers in local governments and those working in non-governmental settings.

I also encourage the Commission to consider recommending any necessary legal revisions 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 (and other) data products. A more explicit acknowledgment that qualified research projects can be conducted for scientific purposes would allow Census to approve studies using confidential data that are primarily designed to replicate existing studies and/or determine the robustness of findings from previous research. Such changes would help ensure the legitimacy of research uses of these data and give greater credibility to the findings based upon these data.

D. ***I encourage greater attention be given to the population representativeness of the policy-relevant research produced using administrative records and population-based surveys to better assess and characterize the population-representativeness of findings from administrative data.*** Many studies use administrative records to "evaluate" the impact of some particular policy or program. As I have argued above, the administrative records provide a potentially cost-effective way of conducting such evaluations—particularly when compared to the alternative of collecting survey data that is collected from a sample representative of the population relevant for the study. But such benefits of using administrative records in evaluative research does not mitigate the importance of assessing the sampling properties of this data source.

Consider the following example 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, PAA and APC 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 its studies.

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 implementing public policy 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.