ChapinHall at the University of Chicago
Policy research that benefits children, families, and their communities

Administrative Data for the Public Good: Opportunities for Advancing Evidence-Based Policymaking using Data Held by the U.S. Census Bureau

A report for the Commission on Evidence-Based Policymaking

Robert Goerge
Leah Gjertson
Elia De La Cruz

2017

Administrative Data for the Public Good: Opportunities for Advancing Evidence-Based Policymaking Using Data held by the U.S. Census Bureau

Robert Goerge
Leah Gjertson
Elia De La Cruz

Recommended Citation

Goerge, R., Gjertson, L., & De La Cruz, E. (2017). *Administrative data for the public good: Opportunities for advancing evidence-based policymaking using data held by the U.S. Census Bureau.* Chicago, IL: Chapin Hall at the University of Chicago.

ISSN: 1097-3125

1313 East 60th Street
Chicago, IL 60637

773-256-5100

www.chapinhall.org

Acknowledgments

We thank Amy O'Hara and Melissa Chiu of the U.S. Census Bureau, Center for Administrative Records Research and Applications (CARRA), for their valuable comments and contributions to this report as well as their ongoing work to support the pilot projects. We thank Kathy Stack, Robin Lipp, and Julie Williams of the Laura and John Arnold Foundation (LJAF) for their guidance and encouragement. We are also grateful to numerous federal agency staff and expert stakeholders who generously shared their time and expertise, providing insightful comments during the development of the RFP and careful review of the individual project proposals.

This report has been published with the generous support of the Laura and John Arnold Foundation. The report is an independent work product of Chapin Hall at the University of Chicago. The views expressed are those of the authors and do not necessarily represent those of the funder.

Executive Summary

To serve the public good, public servants at all levels of government need actionable information to answer questions about what works for the people they serve. Improving secure access to data and information about public program outcomes is essential to enabling policymakers and program administrators at the federal, state, and local levels to engage in evidence-based policymaking–using the best information available to guide decisions and improve government programs.

The use of existing administrative data enables policy-relevant analysis to be conducted more quickly, with greater precision, and at lower cost than alternative methods, making it an important tool in an environment of constrained budgets. Using administrative data can enable statistical analyses for reduced cost and less public burden than if new primary data had to be collected, and using these data sets can enable research that might otherwise be too costly. Further, use of administrative data may provide more coverage, which could allow researchers to generate evidence that more precisely measures the impacts of current programs. Studies using administrative data can often be done faster because data have already been collected, which allows research to inform policy decisions and respond to emerging issues as they occur.[1]

While there are currently secure environments that protect individual program participants' privacy, the current infrastructure within the federal government is not sized properly to meet existing demand from researchers in government or academia to access the federally held data that are needed to rigorously analyze government programs. At the same time, there is increasing recognition that rigorous evidence is needed to make informed decisions about policies and programs. Recent federal policy and legislation reflect the growing importance of using administrative data for policy-relevant program evaluation and research. In March 2016, Congress passed the Evidence-Based Policymaking Commission Act of 2016[2], which created a 15-month-long commission to conduct a comprehensive study on integrating and making administrative data available for these purposes, while ensuring individual privacy and confidentiality.

To address this issue and contribute to the discourse, the U.S. Census Bureau Center for Administrative Records Research and Applications (CARRA) and Chapin Hall at the University of Chicago formed a partnership to demonstrate innovative strategies for combining data across programs and levels of government to advance evidence-based policymaking while adhering strictly to privacy and security regulations. Demonstration projects proposed through this initiative utilize the U.S. Census Bureau Data Linkage Infrastructure, which provides a long-standing, highly secure environment for qualified researchers to analyze de-identified combined program data held by the Census Bureau (see the description of data processes on pp. 1–2). Access to the Census Bureau infrastructure is restricted to qualified researchers who have been authorized through a rigorous screening and approval process and commit to data stewardship protocols.

This report relays findings from this initiative to (1) document demand for administrative and other data sources held by the Census Bureau and (2) inform efforts to design and build a system and processes that facilitate use of federally held data to inform policy making.

---

[1] Office of Management and Budget (OMB). (2016). *Using administrative data and survey data to build evidence*. Retrieved from: https://obamawhitehouse.archives.gov/sites/default/files/omb/mgmt-gpra/using_administrative_and_survey_data_to_build_evidence_0.pdf

[2] Public Law No: 114-140 Evidence-Based Policymaking Commission Act of 2016. Retrieved from: https://www.congress.gov/114/plaws/publ140/PLAW-114publ140.pdf

Demand for Data Held at the Census Bureau

The Request for Proposals (RFP), "Using Linked Data to Advance Evidence-Based Policymaking: Helping Projects Utilize the U.S. Census Bureau Linkage Infrastructure," issued in 2016 by Chapin Hall at the University of Chicago in partnership with the U.S. Census Bureau, solicited research and evaluation proposals to serve as pilot studies that bring in data and demonstrate ways to optimize the Census Bureau infrastructure to provide policy-relevant insights regarding public policies and programs. RFP submissions and experiences implementing the selected projects form the foundation of this report.

The RFP process gathered 45 project proposals submitted by researchers across the country from government agencies including local, state, and federal entities, local and national nonprofit and advocacy organizations, universities, and research organizations (see Appendix A for a list of proposals). Proposals originated in 22 states—Arizona, California, Connecticut, Illinois, Iowa, Kentucky, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, New Jersey, New Mexico, New York, North Carolina, Pennsylvania, Texas, Utah, Virginia, Washington, and Wisconsin—and Washington DC. Projects originating from governmental entities represented agencies at the city, county, state, and federal levels, indicating wide governmental interest in utilizing data sources held at the Census Bureau to inform decision making. The proposals suggested analyses to be conducted with a focus on a variety of geographic areas. Seven projects were national in scope; eight were focused on a specific state; nine were examining county-level data; six were regional or multistate/multicounty projects; six represented city-level projects; and there were two projects focused on schools. Seven projects were proposed with specific study populations.

A broad range of research topics were represented, including education, employment and earnings, health, housing, criminal justice, child support, and taxation and regulation. Some studies involved analyses specific to federal programs like Supplemental Nutrition Assistance Program (SNAP), Temporary Assistance for Needy Families (TANF), and Head Start (see pp. 5–10 for a full list of topics). In their proposals, researchers presented a diverse set of research designs and methodologies, including needs assessments, long-term follow-up of randomized controlled trials (RCTs), mapping, predictive analytics, and descriptive studies to address important policy questions and provide empirical evidence to improve data-driven decisions (see pp. 12–13 for a description of methodology and pp. 10–12 for details on projects' policy relevance).

To be eligible for the RFP, researchers needed to identify one or more sources of "local data," that is, data held by the investigators or their partners that were suitable to be combined with federally held data sources. Examples of local data included research study populations, government administrative data, school records, court records, and integrated local datasets, among others (see pp. 13–15 for descriptions of local data sources). Researchers identified relevant federally held data sources within the Census Bureau infrastructure and requested the Census Bureau produce a de-identified analysis file matching those sources to the local data pursuant to their research design. There was strong demand for earnings and employment data from the IRS and Unemployment Insurance Program, data on mobility and mortality that enables tracking individuals over time and across states, public assistance program data, and the rich information available in Census Bureau survey data (see pp. 15–17 for information about data sources held by the Census Bureau).

Challenges Identified from Efforts to Combine Local Data to Federally-Held Data

Building a system to increase the use of administrative data in the evaluation of public programs requires that we know the challenges as well as the specific needs for accessing and protecting federally held data. From the proposals submitted in response to the RFP, six pilots were selected for immediate implementation (see pp. 6–10 for selected projects). Experiences to date with implementing these projects highlighted barriers in the existing procedures. They also shed light on potential solutions for moving

towards increased usage of federally held administrative data for research and program evaluation while keeping issues of privacy and security paramount. We noted challenges to accessing Census Bureau-held data at both the federal and local level.

We highlight four federal-level challenges for accessing Census Bureau-held data sources (see pp. 19–22 for further details about these challenges).

- Insufficient data documentation: Researchers need better information about data available to develop high-quality proposals. No external metadata viewer or similar documentation is currently available.
- Need to streamline the agreement process: Developing the requisite legal data sharing agreements between researchers and the Census Bureau is time consuming and can introduce barriers.
- Determining who is qualified to access data: Becoming a qualified researcher can be a lengthy process—current guidelines for who and what institutions qualify are subjective and lack clarity.
- Obtaining permission to use data: The Census Bureau negotiates data use conditions with each provider when data are acquired. External researchers then request permission from data providers individually to use their data, a crucial but time-intensive process with significant variation in the length of the approval process and the likelihood of success.

We also found four local-level challenges for accessing Census Bureau-held data sources (see pp. 22–26 for further details about these challenges).

- Establishing data permissions: This pertains to external researchers in two ways. They must ensure compliance with local data sharing agreements as they bring data into the Census Bureau and seek permission from each provider to access existing data sources.
- Accessing data via the Federal Statistical Research Data Centers (FSRDC): All analysis of restricted data must be completed in the protected environment of an FSRDC. Researchers must work in the physical environment in person. However, there are many regions without a nearby FSRDC, and their concentration on university campuses makes them less accessible to government and other nonacademic researchers.
- Transferring data with personally identifiable information (PII) to the Census Bureau: Typically, data providers submit data with direct PII; however, this may not be an option for some data providers. The Census Bureau is exploring alternate options but these have yet to be successfully executed.
- Review by data providers: In addition to the compulsory disclosure review conducted by the Census Bureau, some data providers require review of research products which can include delay of release and requests for additional analysis, raising concerns about censorship of results.

Recommendations for Facilitating Access to Federally Held Data

Despite existing challenges, there is much to be gained by facilitating access to federally held data. Administrative data systems for combining data across sources that are secure and protect privacy greatly increase the ability to generate actionable information about policies at low cost, because they provide richer information on outcomes across domains and give researchers and government officials a single access point to relevant data held by others. However, the full potential of these systems has not yet been realized. We make four recommendations for facilitating access to federally-held data (and provide details on pp. 27–28).

- Invest in increased capacity at the federal level to support projects combining local data with federally held data sources. Specific areas for growth include: developing and maintaining sufficient metadata to support quality proposal development, develop clear and consistent processes for assessing researcher qualifications and obtaining necessary permissions, and ensuring sufficient staff to accommodate the additional work flow.

- Investigate and develop strategies for navigating legal and data sharing issues, streamlining processes whenever possible while also recognizing the importance and need for developing strong and collaborative relationships between all partners and data providers.
- Explore options for cloud-based administrative data research facilities that can increase accessibility while maintaining a high level of privacy and security.
- Plan for sustainability to support the growth of efforts to combine data across sources for the evaluation of public programs and to advance policymaking. This could include exploring options like fee structures and cost-sharing models; having alternate hosts for select sources of federally held data; more direct participation of data providers, such as states, in the Census Bureau Data Linkage Infrastructure; or an administrative data research facility that incorporates federal, state, local, and other public and private data sources.

Table of Contents

Executive Summary i
Introduction 1
Request for Proposals Process and Response 3
Demonstrated Demand for Data Held by the U.S. Census Bureau 5
Description of Proposed Research Projects 5
Local Data Sources 13
Data Sources Held by the Census Bureau 15
Additional Demand for Census Bureau Data Resources 18
Lessons Learned from Efforts to Combine Local and Federally Held Data 19
Federal-Level Challenges 19
Local-Level Challenges 22
Summary of Challenges 26
Conclusion 27
Key Challenges and Learnings 27
Appendix A. Submitted Project Proposals 29
Appendix B. Location of Federal Statistical Research Data Centers 33
Location of Federal Statistical Research Data Centers 33

Introduction

Government policymakers and program managers at the federal, state, and local levels are under increasing pressure to ensure that their programs are achieving optimal results while minimizing costs. In many public programs, policymakers and managers lack secure access to information about program outcomes and to evaluations that can inform policy decisions and strategies for improvement. While there is currently a secure environment that protects individual program participants' privacy, the current infrastructure is not sized properly within the federal government to meet existing demand of researchers in government or academia to access the federally held data that are needed to rigorously analyze government programs.[3] One significant barrier to the availability of evidence is a lack of secure access to key data sets that can reliably measure important outcomes for program participants. In order to design and build a system and processes that meet the needs for accessing and protecting federally held data, we need to understand and clearly articulate the challenges as well as the specific data needs.

To inform efforts to address this problem, the U.S. Census Bureau, Center for Administrative Records Research and Applications (CARRA) and Chapin Hall at the University of Chicago formed a partnership to demonstrate innovative strategies for combining data across programs and levels of government to advance evidence-based policymaking. The goals of the project were to gauge the demand from researchers and state and local governments for having their data combined with other data sources within the Census Bureau's secure infrastructure; demonstrate an efficient way to combine state and local data with Census-held data to answer important questions, while protecting privacy; create compelling use cases for strengthening the Census Bureau Data Linkage Infrastructure to serve multiple levels of government; and inform federal, state, and local strategies for facilitating combining data across programs, ensuring that all strategies adhere to privacy laws and regulations. The Census Bureau infrastructure, which serves as the platform for the demonstration pilots, provides a long-standing, highly secure environment for qualified researchers to analyze de-identified combined program data.[4] The

[3] Program analyses for evidence-based policymaking referred to in this report are used for statistical purposes only (i.e., they are not used for program monitoring or enforcement activities and cannot be used to make decisions concerning the rights, benefits, or privileges of specific individuals).

[4] In the Census Bureau's Data Linkage Infrastructure, de-identified files are those that do not contain direct identifying information on individuals, such as Social Security Number, name, and address. De-identified individual-level files are still restricted use and may only be accessed through the protocols described. In order to remove information from the secure

Census Bureau assigns its own unique linkage keys, Protected Identification Keys (PIKs), to records provided by each project, removes personally identifiable information (PII) from the records, provisions the de-identified analysis file for researcher use in a facility meeting strong IT and physical security requirements, and conducts disclosure review of all output prior to removal from the secure environment. In order to access the Census Bureau infrastructure, all users must be authorized through a rigorous screening and approval process and commit to data stewardship protocols. From the outset, one of the project's explicit purposes was to report experiences and findings to the Commission on Evidence-Based Policymaking.

The primary mechanism for achieving the project goals was the solicitation of project proposals from governmental and academic researchers around the country and the selection of pilot projects for immediate implementation with the CARRA division of the U.S. Census Bureau. The request for proposals (RFP), "Using Linked Data to Advance Evidence-Based Policymaking: Helping Projects Utilize the U.S. Census Bureau Linkage Infrastructure," was issued by Chapin Hall at the University of Chicago in partnership with the U.S. Census Bureau in August of 2016.

This report describes the demand for data sources held by the Census Bureau. The demand comes from governmental and research communities, including local, state, and federal entities; academic and research institutions; and collaborations with nonprofit organizations. It is based primarily on responses to the RFP. It also incorporates additional information on similar efforts occurring concurrently, obtained via institutional knowledge at the Census Bureau and targeted outreach through networks of governmental agencies and academic researchers. This report focuses on data held within the Census Bureau and does not comprehensively represent all existing efforts to utilize federal administrative data for research. In addition, we attempt to put this information in the context of the current use of administrative data for analyzing government programs.

environment, de-identified individual-level data must be fully de-identified (i.e., the data are statistical and sufficiently aggregated so that no individual may be identified).

Request for Proposals Process and Response

The "Using Linked Data to Advance Evidence-Based Policymaking" RFP solicited research and evaluation proposals to serve as pilot studies exploring the long-term outcomes of policies and interventions targeted to inform decision makers regarding public policies and programs. Through the RFP, we sought exciting, ready-to-implement pilots that bring in data and demonstrate ways to optimize the Census Bureau's secure infrastructure to provide policy-relevant insights on federal programs. We also sought development of an inventory of compelling use cases for future projects.

The RFP was released on August 1, 2016 and disseminated widely through public policy and academic research networks.[5] The RFP sought two types of proposal submissions: full proposals for implementation-ready projects and letters of interest (LOI) for future projects that were in a more developmental phase.

The RFP process yielded 45 project proposals (17 full proposals and 28 LOIs). A broad range of research topics was represented, including health, education (from early education through postsecondary), employment, housing, criminal justice, child support, and disaster preparedness. Also proposed were analyses specific to federal programs like the Supplemental Nutrition Assistance Program (SNAP) and Temporary Assistance for Needy Families (TANF). Some proposals focused on special populations such as veterans, immigrants or refugees, and multisystem families. There was also a diverse set of research designs and methodologies, including needs assessments, long-term follow-up of randomized controlled trials (RCTs), quasi-experimental designs, life course trajectory models, mapping, predictive analytics, and descriptive studies. Many of the proposed projects were of very high quality and demonstrate the demand for federally held data as well as the great potential for knowledge generation achieved by facilitating these projects. A list of the submitted proposals is presented in Appendix A.

Full proposals were assessed on seven domains: policy importance and relevance, data and research design, optimizing the Census Bureau infrastructure, high data quality and security for integration

[5] The RFP document can be accessed on the Chapin Hall website at: http://www.chapinhall.org/pages/RFP-Linked-Data-Evidence-Based-Policymaking.

success, research or evaluation expertise, strong commitment of government or nonprofit partner, and the ability to meet project milestones. To be selected, researchers also had to demonstrate that the study met the recognized ethical standards for research with human subjects. The committee that reviewed the proposals was comprised of staff at Chapin Hall, the Census Bureau, and the Laura and John Arnold Foundation, along with expert reviewers selected from federal agencies, including Administration for Children and Families (ACF), Health and Human Services (HHS), Housing and Urban Development (HUD), and the Council of Economic Advisors (CEA). The committee selected six projects to proceed and begin implementation with the Census Bureau. Several other projects identified through this mechanism were not selected for the RFP project cohort but were selected for follow-up by the Census Bureau and federal agencies who expressed strong interest and commitment to the study.

The timeline for submitting proposals and the outreach was done in a somewhat curtailed manner given the urgency to report out quickly. However, given the nature of the response and other communications from those who were interested the RFP but not able to submit a proposal, we believe that the response is representative at least of the range of demand, albeit not the magnitude.

Demonstrated Demand for Data Held by the U.S. Census Bureau

This section provides a comprehensive classification of project proposals across a variety of dimensions, including: submitting organization, topical areas, policy relevance, methodological approaches, types of local data, and types of requested administrative and survey data held by the Census Bureau. The section also incorporates a discussion of key categories of data interest.

Description of Proposed Research Projects

Sources of proposed projects

Projects were submitted by researchers from across the country in government agencies, local and national nonprofit and advocacy organizations, research organizations, and universities. Proposals originated in 22 states and Washington, DC. The 22 states were Arizona, California, Connecticut, Illinois, Iowa, Kentucky, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, New Jersey, New Mexico, New York, North Carolina, Pennsylvania, Texas, Utah, Virginia, Washington, and Wisconsin. Projects originating from governmental entities represented agencies at the city, county, state, and federal level, indicating interest in utilizing data held at the Census Bureau to inform decision making across all levels of government. The proposals suggested analyses to be conducted with a focus on a variety of geographic areas. Seven projects were national in scope; eight were focused on a specific state; nine were examining county-level data; six were regional or multistate/multicounty projects; six represented city-level projects; and there were two projects focused on schools. Seven projects were proposed with specific study populations. Researchers based at universities submitted many proposals; frequently, the academic researchers were collaborating with one or more governmental entities that were seeking to learn from the proposed study.

Study topics

Proposed projects covered a broad range of policy-relevant topics, among which **education, employment, and earnings** were most central. Studies included analyses of how local, state, and federal policies impact educational attainment and earnings outcomes, examinations of postsecondary scholarships and public aid, and long-term follow-ups of experimental studies of employment and early college programs. The descriptions in the boxes for Sample Study 1, “Pathways from K-12 to Work” and

Sample Study 2, "Effects of Public Need-Based Aid for College," illustrate two projects focused on education, employment, and earnings. Other education and employment studies encompassed specific population segments such as veterans, students with disabilities, minority students, and low-income/disadvantaged urban and rural schools.

Sample Study 1. *Pathways from K-12 to Work*

This project characterizes students' pathways from K–12 schooling to work, with particular attention to those students whose trajectories are negatively affected by failure to graduate from high school or fail to engage in work or school after graduation. These are the Opportunity Youth that are of so great concern at all levels of government now. The study uses a longitudinal cohort of students from a large, urban school district with linked demographic information, school achievement test scores, school attendance, high school graduation, college attendance and graduation, and juvenile and criminal justice involvement. They intend to match this dataset through the Census Bureau infrastructure to individual and business tax data from the IRS, employment and earnings data from state UI wage records, data on receipt of Supplemental Security Income (SSI), and mortality.

The resulting combined data can answer the question "what early life educational and out-of-school experiences predict low earnings, unemployment, not attending post-secondary education, and incarceration?" Results of this study can provide actionable results for both schools and the postsecondary institutions, be it employment or education.

Sample Study 2. *Effects of Public Need-Based Aid for College*

This project augments an existing large, longitudinal state database by adding income and earnings measures to measures of college enrollment and eligibility for college financial aid. The core of the project is an evaluation of the impacts of a state program that lowers the price of college for low-income residents.

The original database consists of all state residents applying to receive state and federal financial aid at state colleges and universities, during the school years 2007–08 through 2014–15. They observe college outcomes for applicants and request matching to income and earnings from IRS tax records and state UI wage data. An additional match to the American Community Survey will provide a fuller picture of education and employment for a subset of applicants. The study is poised to answer questions about the longer-outcomes of the public financial aid program and inform future state investments.

Several studies focused specifically on **early care and education**. Proposed projects about early care and education took multiple forms. Some focused on the capacity of such programs to improve the lives of participating children and their families in the short term and over the life course. Other studies aimed to improve enrollment, develop effective program assessment tools at the local level, and examine policy dimensions of the federal child care subsidy programs.

Another common theme was studying **public assistance programs** such as Temporary Assistance for Needy Families (TANF), the Supplemental Nutrition Assistance Program (SNAP), and state and local social services. Multiple projects emphasized needs assessments and eligibility determination at the state and local level. Several of these proposed projects incorporated a geographic or mapping component. Similar research, typically presented by state or local governments, was focused on the identification and service utilization patterns of high-need or multisystem families. This research was designed to inform decisions around the strategic allocation of public resources as well as identify areas of need. Finally, researchers submitted plans for follow-up studies of public assistance program participants to build on existing experimental study populations to examine long-term outcomes.

Housing and homelessness was a prominent theme among submitted proposals. Researchers aimed to study the consequences of eviction and the service use trajectories and outcomes of homeless families. The descriptions in the boxes for Sample Study 3, "Does Eviction Cause Poverty," and Study 4, "Service Utilization of Families Experiencing Homelessness," detail projects on these topics. Also of interest, particularly at the local level, was establishing correlations between housing stability, child school attendance, and educational outcomes. Several studies were focused on the use of public housing subsidies to determine the impact of program participation on household economic status (wealth, debt, and poverty), among other indicators of family well-being.

Sample Study 3. *Does Eviction Cause Poverty?*

This project evaluates the causal impact of eviction on employment and schooling outcomes in a large urban county using a database containing the universe of eviction case court records from 2000–2015. The research design leverages the random assignment of eviction court cases to judges, which creates a natural experiment for studying the causal effect of eviction on a wide range of short- and long-run household outcomes associated with poverty. They propose combining court records data with Census Bureau held data to study earnings, employment, receipt of TANF and SNAP, and homelessness for evicted adults. Using school district data, they also propose to study how eviction affects chronic absenteeism, misconduct, and performance on standardized tests among children of evicted households.

Results will inform the policy debate on housing assistance, by highlighting the costs of eviction, and by providing solid evidence on the long-run impact of eviction on families. This study will contribute towards the debate on tenant-based versus place-based housing assistance.

Sample Study 4. *Service Utilization of Families Experiencing Homelessness*

The goal of this study is to provide evidence to the experiences of families experiencing homelessness to inform policy efforts to better serve this population. The project uses a linked database of social services, school records, and law enforcement records for one large county. They intend to use the Census Bureau infrastructure to 1) develop a comparison group using the American Community Survey, and 2) link to federal program data including HUD housing data and change of address data.

This data linkage will allow the researchers to longitudinally examine housing patterns, compare these patterns among families who were housed and those who were homeless, and to identify risk and resilience factors for subsequent homelessness, child welfare involvement, criminal justice involvement, and child academic and behavioral outcomes.

Another common theme researchers wished to pursue was **health**. Researchers submitting proposals for health-related studies wanted to examine a variety of policy-relevant topics across the life course, including prenatal and postnatal care, childhood lead exposure, and suicide and predictors of mortality. The proposed studies aimed to investigate the predictors and consequences of these health events, using policy variation in public health spending and the availability of programs over time, as well as aggregate data on indicators such as health behaviors and income inequality. The description in the box Sample Study 5, “Health at Birth and Later Life Outcomes,” details an example of a health-focused project.

Sample Study 5. *Health at Birth and Later Life Outcomes*

This study aims to document the relationship between health at birth and later life outcomes, while evaluating the returns to policy-driven early health investments. First, they will use within-family variation in birth weight to estimate the effect of health at birth on adult health and productivity. Second, they will analyze two specific public health investments that target the prenatal and neonatal periods for low-income pregnant women in one state. They propose combining birth records for over 25 million individuals born in one large state over six decades with IRS data and Census-held administrative data to examine income, educational attainment, health, and use of government services later in life.

This project analyzes the role of policy interventions in promoting lifelong improvements in health and productivity. Given the many large social programs that exist in the U.S. to promote the health of pregnant women and infants, additional information on the relationship between public policy, health at birth, and later life outcomes will add substantial value to ongoing policy debates.

In the remaining handful of proposals, researchers touched on other topics of policy salience. Two researchers proposed to study immigrant and refugee populations, exploring the effectiveness of resettlement interventions on employment, self-sufficiency, and other economic outcomes. Several other researchers aimed to examine criminal justice topics, including patterns of recidivism among youthful offenders and impacts of parental incarceration on children and families (see box for Sample Study 6, "Program Utilization by Formerly Criminalized Youth"). Other projects also focused on family issues, such as child custody and child support arrangements, attempting to isolate the effects of legal practices and state policies on affected children and parents, and to improve the robustness of analyses through the use of richer data. One researcher aimed to improve public safety response in natural and manmade disasters with predictive models of demand for services in emergencies. A final category of studies included research on the impacts of taxation and federal regulations on economic outcomes for individuals, firms, and the broader economy.

Sample Study 6. *Program Utilization by Formerly Criminalized Youth*

This project proposes to combine state agency data on juvenile justice, criminal justice, and recidivism with Census Bureau-held data with respect to young persons who have been involved in criminal or juvenile justice in one state. The study brings juvenile recidivism data integrated from three state agencies for nearly 83,000 individual juveniles who were involved in juvenile or criminal justice between 2000 and 2014 with information about arrests, charges, disposition, sentencing, incarceration, and rearrest. This project seeks matching with federally held data resources on income and employment (IRS and state UI wage data), housing assistance (HUD), health insurance and health care (Medicare and Medicaid), and location.

By combining state juvenile recidivism data with data resources held at the Census Bureau, the study aims to identify government programs whose utilization is associated with non-recidivism, higher incomes, better health, more stable housing, or other socially desirable outcomes. Results of this study can contribute to policy discourse about how to serve young offenders and identifying strategies for reducing recidivism.

Policy Relevance of Proposed Projects

In their proposed studies, researchers presented policy salience across multiple dimensions, providing policymakers with empirical evidence to consider changes in programs or policies, improve data-driven policy decisions, and inform decisions about governmental resource allocation. Access to administrative data from multiple federal and state programs will help an evaluator better understand the net cost of an intervention. A behavior health intervention might increase or decrease health care usage. A homelessness intervention could improve employment rates and decrease reliance on TANF or SNAP.

Proposal research questions are contextualized in current policy changes, providing frameworks to analyze the as yet unmeasured effects of policies. One brief example of a policy change study is a project examining state-level variation in minimum wage legislation. Some researchers make the case for policies that could be improved once data show that allocation of resources or program eligibility is not handled accurately. An example is a proposed study of the income threshold for Head Start eligibility that explores the impact of adjustments for cost of living and other factors on which families are eligible for the program. Heading into the reauthorization of the Head Start Act, it is important to understand how the eligibility requirements are functioning. In addition, measuring the appropriateness of requirements for program eligibility more broadly is of increasing relevance when changes on minimum wage legislation take place in different states. This is a topic that could be of interest to state and local governments as well as the federal government. In their proposals, researchers demonstrate policy salience across multiple

dimensions, which are illustrated by specific examples of pilot projects selected from the responses to the RFP.

Access to the federally held administrative data sources provides researchers with information that greatly enhances the policy relevance of their work by enabling analyses that would otherwise be infeasible or restricted by inferior data quality. For example, tax records from the IRS provide a reliable source of earnings information for individuals over time for the full universe of tax filers. Without IRS records, researchers may have no other means of accessing earnings data or may be limited to self-reported income information from a subset of respondents in certain time periods. The value of tax records for evaluating public programs and contributing to the policy discourse is clearly displayed by the recent work of Raj Chetty and colleagues[6], among others. Pilot studies that illustrate the potential gains of using IRS tax records for evidence building in education are detailed in Sample Study 1, "Pathways from K-12 to Work," and Sample Study 2, "Effects of Public Need-Based Aid for College." Both studies use tax information to improve data-driven policy decisions surrounding K–12 and postsecondary education.

Sample Study 5, "Health at Birth and Later Life Outcomes," also seeks tax records combined with other sources of administrative data to inform governmental decision making by providing insights into the long-term benefits of policy investments in prenatal and infant health.

State and local agencies may seek to conduct resource allocation or needs assessments by aggregating data between and combining information from agencies and sharing it among agencies at the county or state level. These combined databases, supplemented with the data resources available through the Census Bureau infrastructure, create opportunities to effectively identify and respond to emerging trends and issues and inform decisions about governmental resource allocation. Sample Study 4, "Service Utilization of Families Experiencing Homelessness," and Sample Study 6, "Program Utilization by Formerly Criminalized Youth," provide examples of projects that use federally held data resources to maximize the potential for knowledge generation from combining data across sources at the state and local level.

Researchers may also use access to federally held data to generate empirical evidence that informs consideration of changes in programs or policies by addressing larger questions of poverty and self-sufficiency and the interactions between law, policy, and government support systems. One example of this type of research is Sample Study 3, "Does Eviction Cause Poverty?", which uses a quasi-experimental design to examine the causal impacts of eviction on subsequent economic and family outcomes, assessing the effect of an eviction event on poverty trajectories.

[6] A list of papers by Chetty and colleagues is available here: http://www.rajchetty.com/papers-chronological/.

Finally, connection to federally held data sources presents practical advantages such as cost and time savings. Matching data from participants of a study, held by PIs, to administrative and survey data held by the Census Bureau could be cost effective in two ways. First, enriched datasets could allow a more precise measurement of program impacts, providing a deeper understanding of the effects of programs designed to increase self-sufficiency for low-income and disadvantaged populations. Second, using data to identify where resources are more effective could improve the use of public funds.

Methodology and Research Design

The overall goals of the submitted proposals can be largely grouped as follows: those that aimed to reach a better characterization of the vulnerable population that uses government programs, those that aimed to improve service delivery, and those that aimed to measure long-term policy impacts. Strong proposals linked the goal, methods, and policy relevance to the data sources. For instance, researchers submitting proposals with the goal of improving the allocation of resources suggested the use of spatial/geographical analysis or, more specifically, GIS mapping and use of census tracts. As open-source coding and user-friendly software become more widely available, geographical analyses provide a way of mapping eligibility and demand for services and maximizing limited government resources. Geographical analyses need rich demographic, income, and other information to successfully depict where demand should meet supply of services. There are interesting cases where comparing Census-held data to PI-held data would be useful for examining how measurement affects statistical inference, such as studies that aimed to compare income or child support reported in surveys against actual income from IRS forms. Comparing both measures can shed light on the magnitude of misreporting, which is a valid concern among these researchers since misreporting affects the understanding of a program's impact. In some cases, researchers sought information needed for mapping that can already be accessed through public sources (i.e., surveys or aggregate-level indicators), which suggests there is not universal awareness in the research community about the Census Bureau's existing public use datasets.

In another set of proposals, researchers aimed to exploit unique access to databases containing baseline and follow-up information of randomized controlled trials (RCTs) that have already taken place. In these proposals, researchers' main empirical method is simple mean comparison. In these cases, random allocation of participants eliminates selection bias and the difference in means can be attributed to participation in a program. The advantage of combining databases from RCTs is that causal inference of employment, housing, and education programs can be expanded beyond the intended realm of the experiment and can investigate long-term effects many years after the intervention. Spillovers to other domains like employment, welfare, economic mobility, and other long-term outcomes is promising. The

main drawback is that not all researchers may have consent from original participants/beneficiaries of the randomized allocation to combine their baseline information with administrative data.

Some researchers submitted proposals incorporating research designs that exploit existing variation in administrative processes, program implementation, time, and geography to create opportunities for causal explanations. Examples of these methods include: (1) using judicial assignment to predict outcomes following a court event; (2) examining participants just above or below a cutoff for receiving service; and (3) variation created by staggered implementation of a statewide evaluation of program impacts. These methods do not replicate an experiment but can improve the rigor of research findings within the context of existing systems and datasets. They can be implemented in circumstances when an RCT may be infeasible due to ethical or cost constraints.

In a few proposals, researchers mentioned predictive modeling and machine learning for text mining. These sophisticated methods are useful for estimating long-term trajectories of a wide range of outcomes. The computational techniques can also be useful for analyzing court records, legislation, and other inputs that are not presented in a traditional data format.

Local Data Sources

In their project proposals, all researchers were required to identify one or more sources of "local data." Local data are defined as data held by the proposal authors or their partners that are suitable for combination with data held by the Census Bureau. Local data necessarily included some PII in order for the Census Bureau to assign its PIK linkage keys, even though researchers access de-identified files within the secure environment. Local data sources took a wide variety of forms, which are presented here categorically.

Government administrative data

One source of local data was government administrative data. Researchers proposed projects to supply data from county, state, and federal entities and submit it to CARRA to be combined with other data sources. Government administrative data included public benefits programs, state or county records of federally funded programs (e.g., SNAP, TANF, Medicaid, WIC, and EITC), vital records, health records, child welfare, child support, and data from social service providers. In many instances, proposal authors had access to data that had already been integrated across sources at the local level. One example of integrated data is the use of information combined across law enforcement and criminal justice agencies to track offender recidivism at the state level. Another example, also at the state level, combined data from state public aid programs with postsecondary education records and wage records to observe education and earnings attainment. Projects of this nature can enrich already robust data with federal

information. Key gains from the federal sources include the capacity to track individuals as they cross out of the county or state at issue and to obtain wage and earnings information and public program participation that is not available at the local level.

School records

School records from early education through postsecondary education represent another category of local data in the project proposals. School data was sourced by school districts, state or local governments, and regional or national agencies. The information in school records varied across projects and included test scores, special education, high school graduation, postsecondary education enrollment and completion, school fiscal information, and attendance, teacher, and staff records. In general, having school records enables research on a broad range of education topics and allows for the rigorous evaluation of specific policies.

Court records

Court records provided the local sample data for several project proposals. Using court record data provides a strategy for identifying populations that have experienced a particular event, such as eviction, mortgage foreclosure, bankruptcy, or child custody orders. Combining such individual data with federally held data at the Census Bureau generates the opportunity to observe outcomes over time that predate and postdate the court event. Court record data are relatively accessible via web scraping techniques or by purchase from the state or county. The availability of court records over time is constrained by when a particular jurisdiction transitioned to electronic records. Court records typically provide name and address data at one point in time. Limited information still allows individual records to be combined across sources, although the match is of lower quality than data identified with additional personally identifiable information (PII) such as social security number (SSN) and date of birth.

Contemporary and previously existing study populations

In this category, the proposal authors are executing research studies or have access to data from previously conducted studies, often of experimental design. Study populations for a couple of proposals were comprised of samples from active or enrolling projects. These researchers are looking to use data sources at the Census Bureau to complement the program or survey data collection already occurring. More commonly, however, researchers proposed to use prior study populations that were generated one or more decades ago. These researchers typically focused on long-term follow-ups of participants that would be impractical without access to federally held data. Such projects leverage previous research investments and provide much sought-after information about the long-term impacts of policies and programs at a relatively modest cost.

Publicly available aggregate data

A final source of local data emerging from the proposals is aggregate data that is publicly available. Examples of these include information on natural disasters, public safety, and government regulations. Such data sources may be readily available online in a complete format or may require significant time and effort to compile from across many sources. The most common use for these data are in combination with other data sources to provide contextual or ecological information. These data do not typically contain individual records and may be integrated with individual-level data by geographic location (e.g., county, zip code, census tract).

Data Sources Held by the Census Bureau

Each researcher requested their local data sources be combined with one or more sources of data currently held within the Census Bureau infrastructure. All data holdings were assigned the Census Bureau's PIK linkage keys so that researchers accessed de-identified files within the secure environment. The Census Bureau worked with selected proposals to ensure data minimization (i.e., that all requested data were relevant and necessary for the analysis). The administrative data sources were primarily sourced from federal agencies but there were also some state and third party data.

IRS records

There is strong demand for access and data matching to tax records from the Internal Revenue Service (IRS). Tax information includes individual and earnings information from forms such as the 1040, 1099, and W-2. Using tax record data enables researchers to observe filers' earnings over time and across state borders. This addresses a key limitation of much of the work done with state-level administrative data. For example, if researchers are investigating the return to education at the high school or postsecondary level, they would ideally observe the earnings histories of all graduates 10 years post graduation. They may have administrative data on wages from the home state but lack data on anyone that has moved out of state. This presents a particular issue for large metropolitan areas with economic and residential activity that extends into multiple states. Another benefit of using tax records is the availability of marital status and name changes to improve record matching over time.

Researchers also want to access tax records on filing and receipt of the Earned Income Tax Credit (EITC), the refundable tax credit that supplements employment wages among low-income working families.

Wages and employment

Another sought-after source of earnings and employment information is Unemployment Insurance (UI) wage data. UI data are typically generated on a quarterly basis and provide information on wages and spells of unemployment by quarter. These data, collected and managed at the state level, complement

earnings information from tax records which captures some types of earnings, such as self-employment, that are not available in the UI wage data. Because UI wages are managed by the states, each state sets the parameters around whether and to what extent their data are available for combination with other data sources and research by external users. Some states allow external researchers to use their UI data through the Local Employment Dynamics program at the Census Bureau. Access is given after a case-by-case review of the research request. A minority of states make their data available to external researchers through the Center for Economic Studies research network. Another group of states prohibits any external data use. This circumstance creates a patchwork of availability for UI wage data.

Mobility and mortality

Federally held data at the Census Bureau offers the opportunity to do significantly better longitudinal research than what is typically possible with siloed state administrative data. The Census Bureau holds data from multiple, or all, states and individuals' status can be tracked over time. In order to track outcomes over time, researchers need to observe where individuals are located throughout the life course, including where and when they die. Federally held data sources of address changes are a key tool for observing mobility, especially when people cross state lines. These moves cannot be observed in state-level administrative records. Death records are also important for research on program and policy interventions. Death records can indicate who in the local data is still eligible for observation. In addition, mortality is an outcome of interest in some studies.

Census Bureau

There was also significant interest in access to restricted versions of Census Bureau data, including the decennial census, the American Community Survey (ACS), and the Current Population Survey (CPS). The restricted data allow for individual records to be combined between local data sources and the rich information collected through the Census Bureau surveys. The decennial census,[7] conducted every 10 years, is the official counting of the American populace which is used to apportion the House of Representatives and inform many other aspects of governmental decision making and resource allocation. The decennial census data has the largest pool of individual records available for matching; however, it is only collected every 10 years, and only the 1940, 2000, and 2010 records are currently available for electronic matching. The decennial census provides one strategy for filling in demographic information that may be missing in administrative data. It also provides a method of looking at household structure

[7] Decennial Census of Population and Housing. https://www.census.gov/programs-surveys/decennial-census.html

and offers the opportunity to identify multiple generations for studies looking at intergenerational outcomes.

The ACS,[8] fully implemented in 2005, replaced the long form of the decennial census with smaller monthly samples of the American population. It provides information on population demographics, housing, family and social characteristics, education, and information on employment, earnings, and economic status. The CPS is also collected via a monthly sample with items focused on labor force participation.[9] In addition to the regular CPS questions, there are periodic supplemental questions on a variety of topics, such as child support, health insurance coverage, and school enrollment. The ACS and CPS are collected for small subsamples of the population, meaning that research projects need large samples of local data in order to potentially benefit from matching data at the individual level. However, the sampled data may also be used to generate a comparison group for administrative records, if there is a model or definition for determining the comparison group within the survey. For the pilot projects, the Census Bureau data sources may be the first data available for matching because there is not the additional process of receiving approval for data use from an external provider agency.

Public assistance

Combining local data samples' public assistance records creates opportunities for observing program utilization across programs and over time. It also provides information on the results of experimental interventions, responses to policy changes, and time trends. Many projects requested access to federally held data sources of public assistance receipt for programs like Medicaid, Medicare, Supplemental Security Income (SSI), housing assistance, and child care subsidy programs. Subsets of these administrative databases are submitted to CARRA from federal agencies like the Department of Health and Human Services (HHS), the Department of Housing and Urban Development (HUD), and the Social Security Administration (SSA). CARRA also holds TANF and SNAP data; however, these data are provided by the states rather than federal agencies. There was significant demand in the project proposals for matching to TANF and SNAP data. However, the years and jurisdictions of available TANF and SNAP data varies. Almost all states submit TANF data to the federal government but almost half the states only provide a sample of their TANF population. Therefore, data from those states cannot be combined with the universe of program recipients. SNAP data are only available for a subset of states that have shared that data with CARRA.

[8] American Community Survey (ACS). https://www.census.gov/programs-surveys/acs/about.html
[9] Current Population Survey (CPS). https://www.census.gov/programs-surveys/cps/about.html

Additional Demand for Census Bureau Data Resources

The Census Bureau is also experiencing interest in its Data Linkage Infrastructure outside of this RFP process. For instance, HUD has partnered with the Census Bureau to host two important randomized evaluations, Moving to Opportunity and the Family Options Study. These evaluations are being hosted for broad use in analyzing long-term outcomes of public programs, public policy, and demographic, economic, or social conditions. Researchers applying for access to the data may propose data matches and evaluations and can even bring in their own data for combination with other data sources. HUD's partnership with the Census Bureau addresses three issues: (1) HUD's lack of capacity to continue to issue data licenses for requests for restricted data access; (2) the Census Bureau's strong data security infrastructure and long-standing record of data stewardship protection, when used to host HUD data, made it unnecessary for HUD to maintain duplicate systems to ensure security of data requests; and (3) HUD desired increased research capacity for studying knowledge gaps and new, previously unanswerable, research questions leveraging data resources at the Census Bureau. This partnership will produce a line of research on housing interventions and related issues.

Further, the Census Bureau has commenced numerous evidence-based pilot projects, particularly at the federal or national level. One such project is being conducted with the Office of Economic and Manpower Analysis (OEMA) of the U.S. Army. This research will combine Army administrative record data to study the labor market outcomes and socioeconomic well-being of Army applicants, service members (including officers, warrant officers, noncommissioned officers, and soldiers), and their families during and after military service. In particular, the study will examine the effects of military relocations on family outcomes; the effects of military service experiences on employment, earnings, and risk of suicide; and the long-term effects of parental absences on children. Another project, with the National Association of Manufacturers (NAM), will examine the labor market outcomes of people who received a manufacturing-related credential or were enrolled in such a program. The study will combine data from two private sources—the NAM individual file and the National Student Clearinghouse—with Census-held administrative data. These projects are just a few examples of recent interest in utilizing the Census Bureau's infrastructure for evidence-based program research and evaluation.

Lessons Learned from Efforts to Combine Local and Federally Held Data

Building a system that meets the needs for federally held data requires that we know the challenges as well as the specific needs for data management and security. This section describes what must be done to increase the use of administrative data held at the Census Bureau in the evaluation of federal, state, and local programs, based on the experiences to date implementing the pilot projects.

Federal-Level Challenges

Insufficient data documentation

Researchers need better information about the data that is available. Some proposed projects were not feasible. Bad design was not the primary reason for this lack of feasibility. Rather, there was a lack of detailed knowledge of the data sources held at the Census Bureau, resulting in research plans that could not be implemented due to issues such as data structures, availability of specific variables, and conditions of the sample that the researcher was unable to observe in advance. The Census Bureau currently has no metadata viewer available to aid researchers in developing high-quality research proposals. Clearly, there is a shared responsibility to provide this in the future. Better metadata is needed, preferably metadata that describes the provenance, contents, quality, and prior analyses of the data. This is one of the most common recommendations that has been made since the beginning of the use of administrative data for social science purposes. However, little progress has been made. Researchers working with this type of administrative data have found it difficult to understand the qualities of the data available, which limits their ability to develop a quality research design and reduces confidence in understanding and interpreting results.

The responsibility for this partly lies with the data providers/stewards/curators and users. Since the administrative data that CARRA receives is not generated internally, but received from federal or state agencies, the responsibility for better metadata begins where the data are generated—at the local level. However, given the uneven documentation available from the agencies that provide the data, the creation of metadata is usually left to those who have an incentive to make the data easier to use. Improving

metadata requires that users have the experience and capacity to analyze the contents and quality of a particular dataset and use that analysis to decide whether the data are acceptable for their specific research task.

Important issues that should be considered include the comprehensiveness of metadata that is required and the appropriate platform for making metadata available to potential researchers. One task would be to determine the appropriate level of metadata that enables researchers to design and propose quality studies while protecting restricted data and not placing an undue burden on the data providers or the Census Bureau staff. For example, variable lists and value labels provide some information. However, those sources would be more detailed if they contained descriptive information about each variable, such as the minimum, maximum, and mean values as well as frequencies and the rate of missing data. Providing information on the coverage of a specific population that the data provides is important. Richer still would be a platform that allowed for subject matter experts (SMEs) and data users to provide feedback on known limitations and solutions.

A related issue is the budget and staffing capacity of the Census Bureau to develop, maintain, and disseminate metadata. Are there dedicated staff available to answer questions about the data sources? How is the metadata shared with potential researchers? Is it published on a public-facing webpage maintained by the Census Bureau? Are there any restrictions or requirements for who can access and navigate the information? Is some metadata available publicly while more detailed information can only be accessed once researchers are using restricted data within the Federal Statistical Research Data Center? The ability to improve and disseminate metadata in a scalable and sustainable way, such as in the context of an administrative data research facility, would likely require a capital investment.

Efforts to streamline the agreement process between researchers and the Census Bureau

This represents another well-known problem around increased use of administrative data—legal agreements that are acceptable to lawyers and researchers that do not insert additional, time-consuming barriers into the process. This agreement is often what substitutes for a long-standing trusting relationship between data providers and researchers. Data providers need to have a sense of comfort that their data is protected and not used in a manner that they do not approve. As part of the pilot project initiative, the Census Bureau developed a new template to streamline the process for executing legal data sharing agreements between project PIs and the Census Bureau. It is unclear if this new template form and process will suffice to meet legal requirements and be put into practice for future projects. States also have capacity issues, with limited staff available to address these issues.

If the new template is not ultimately approved for use, then the pilot project researchers will have to develop agreements through the existing channel, a process which can take up to a year to execute. The

typical, customized agreement process begins with conversations between programmatic representatives on the research, the data, and data uses. Once an agreement is drafted, there are multiple layers of legal review: the Census Bureau, the Department of Commerce, and the PIs and their organization. Agreements using Department-approved templates, however, do not require additional review by the Department of Commerce. Whether using templates or customized agreements, it will also be necessary for the Census Bureau to increase capacity to take on additional projects.

A general recommendation here is that there needs to be better training of lawyers to ensure agreements are reached faster. Templates are one strategy to address this, but given the need to address specifics around each particular agreement between two (or sometimes three) parties, templates can be more of a barrier. For instance, templates often require information to be shoehorned into the agreement when it is easier to create a separate agreement that includes all of the concerns of a local agency. Templates may raise issues that are not of central importance, but they are often difficult to dismiss once stated.

A key aspect of the outcome of any agreement has to be transparency around what is happening to the data during analysis (who is accessing it and scope of use) and to the results of the research once it is done. These components must be crystal clear in order for data providers to have the comport they need.

Determining who is qualified to access data resources held by the Census Bureau

Projects using Census Bureau-held data must comply with underlying data sharing agreements between the data provider and the Census Bureau, as well as U.S.C. Title 13. Agreements not only describe how the Census Bureau handles data, its security systems for storage and processing, and data disposition, but they also dictate uses of the data (PII versus de-identified files) within the Census Bureau infrastructure and any procedures for informing data owners of products using their data. A small number of Census Bureau employees have access to identified files for data processing purposes and researchers are allowed access only to de-identified files within the secure environment. While some data owners also restrict data users to a specific group of people, the vast majority of data users are defined as "qualified researchers" who are Census Bureau employees or researchers with Special Sworn Status. After a review of other data provisioning models in the federal government, the Census Bureau has developed criteria for reviewing evidence-building project proposals, including who is a qualified researcher. Evidence-building project reviewers will consider a principal investigator's (PI) affiliation—such as a research organization or a research unit in a larger organization—and the affiliated organization's mission. Researchers with other affiliations may demonstrate their expertise and experience to conduct and complete the proposed project. Finally, proposed projects must be statistical in nature, with a research or evaluation purpose and not a direct enforcement function or any other purpose that might affect a specific individual's rights, benefits, or privileges.

Obtaining permissions to use data sources for external research

Most of the administrative data sources held by the Census Bureau, other than the Census surveys, are sourced from other federal agencies, state agencies, and a few third party providers. In order for these data to be used for external research, permission must be obtained from each data provider. This issue is directly relevant at the federal and local level. At the federal level, the Census Bureau negotiates the conditions of use with each data provider before data are transferred to the Census Bureau infrastructure. Conditions to be negotiated include, are the data available only to the provider? Can the Census Bureau use the data? Can the data be used by other researchers, and if so, does the data provider make this decision on a case-by-case basis?

At the local level, researchers must make these decisions for the data they bring to the Census Bureau, often in collaboration with their local data partners (such as one or more state or county agencies or school districts). Executing all the necessary agreements is often a complex task. The data sharing requirements and status of preexisting agreements, such as those between multiple state agencies, can vary dramatically, requiring a time-intensive individualized approach.

In negotiations with data providers, it is key that each provider perceives a tangible benefit from the results of combining their data with other sources. This can make them a more willing partner and smooth the path through the required legal processes. The strength of the relationships between the data providers and researchers is often a crucial factor in the successful navigation of this process. Moving forward, one potential solution to incentivize local data providers to granting broader data use permission is to offer lower cost matching to the federally held data sources in exchange for making their own de-identified data available for use by others. This will help to build the repository of administrative data available for research in the secured Census Bureau environment.

Local-Level Challenges

Obtaining permission to access data held at the Census Bureau

In order to access data held by the Census Bureau, researchers must obtain Special Sworn Status (SSS). They are required to undergo an extensive screening process that typically takes months to complete. There should be no shortcuts in the process of certifying an individual researcher to access restricted data. In some cases, the process may take longer because there is information that raises concerns (sometimes due to mistaken identity). However, there is also a capacity issue in those Census Bureau units that have the responsibility for processing and screening applicants for SSS. Particularly as the 2020 Decennial Census nears, there is a danger that the demand for SSS will overwhelm the Bureau's capacity to provide it.

While the SSS review process itself takes 4 to 6 weeks, the whole process of becoming qualified, beginning with application, often takes longer, depending on several factors. Most eligibility requirements are known and documented, such as non-US citizens must have resided in the United States for three of the last five years. However, other eligibility requirements vary and are relatively unknown and not documented. In order to become eligible to become an SSS employee of the U.S. Census Bureau, each person must submit an onboarding application. When an application is reviewed, specific information will be requested from the individual based on the security level being sought. The duration of the background check varies based on whether there was a previous investigation completed and how long the applicant takes to respond to inquiries. The entire process could take several months. Once all requirements have been met, a notice will be sent to the requestor informing them of their SSS with the Census Bureau. As a condition of SSS, the researcher must adhere to all data stewardship protocols for handling data within the secure environment.

Compliance with local-level data sharing agreements

It is frequently necessary to modify existing agreements or create new agreements with local data providers (e.g., state or county agencies) so that data can be shared with the Census Bureau. While the capacity to do this varies by local area, most often this will take many months. Local data providers need to learn about the organization to which the data are going (CARRA) and understand the legalities of sending data to a new federal agency, particularly when doing so for the first time. Local entity concerns about ensuring compliance with existing data protection laws, such as the Family Educational Rights and Privacy Act (FERPA, for education data) or the Health Insurance Portability and Accountability Act (HIPPA, for health data), require navigating additional legal issues. Even with CARRA staff providing agreement examples, templates, and other documents to assure data security and continued local agency control, any new arrangement can take months, or even years, to fully execute.

A related issue is the extent to which local data holders are able to obtain permissions and appropriate human subjects consent for the use of identifying information from existing studies or surveys for matching to federally held data sources. The determination of what is feasible will be made on a case-by-case basis and will depend on what study participants agreed to when consenting to participate in the original study. For studies using existing data study and survey populations, especially for studies that took place years or decades ago, going back to participants to obtain consent for new research would likely be impractical and cost prohibitive. This challenge could be addressed by surveys and studies rethinking time limits for data usage when consent is originally obtained from participants. Forward-looking research agreements that enabled long-term follow-up would need to address issues such as plans

for maintaining identifying information and receiving consent to track participants over time using administrative data.

Accessing data via the Federal Statistical Research Data Centers

All analysis of restricted data must be completed within the protected environment of a Federal Statistical Research Data Center (FSRDC), physical locations that provide highly secure IT environments for access to sensitive data housed on a secure server located elsewhere. Researchers must go in person and conduct all data analysis in that environment. While a few of the selected proposals have ready access to an FSRDC, others will need to travel to the nearest FSRDC.

There are currently 24 FSRDC facilities in locations around the country. These facilities are partnered with universities, nonprofit research institutions, and government entities. Although many areas of the country have FSRDC sites (see Appendix B), there are many regions where a researcher would need to travel a significant distance to access an FSRDC site. Further, these sites are concentrated on university campuses and may be less accessible to government researchers. The need to analyze Census Bureau-held data at an FSRDC is a known obstacle that needs to be addressed.

FSRDCs receive funding from the Census Bureau but are locally administered. They are responsible for covering a certain percentage of operating expenses. Once the pilot projects were selected through the RFP, PIs identified the FSRDC closest to their organization and the Census Bureau reached out to them for cost estimates for the required work. Because the FSRDCs have not previously been used in this fashion, getting cost estimates was a lengthy process. The pilot projects received a wide range of prices, depending on the selected FSRDC site. As a result, the cost of a given project may vary dramatically, depending on the research team's distance from the nearest FSRDC. Some FSRDCs may also have limited capacity to quickly develop budget estimates for new projects. These circumstances present additional barriers for researchers seeking funding for research projects and needing to develop accurate budgets.

One solution to the difficulties presented by the existing FSRDC infrastructure would be a cloud-based platform that retains all the necessary security measures while providing researchers more flexibility with regard to physical location. Exploring the feasibility of such a platform to facilitate research on government programs is one of the stated goals of the Commission.[10] The Census Bureau is investigating the use of a truly scalable cloud-based government data research facility via an informational pilot, the Administrative Data Research Facility (ADRF) project.

---

[10] Evidence-Based Policymaking Commission Act of 2016, Pub. L. 114-140, 130 Stat. 317. https://www.cep.gov/content/dam/cep/about/public-law-114-140.pdf

The ADRF is being developed and deployed through a partnership with New York University and the University of Chicago. The ADRF system architecture is comprised of a secure computing environment for processing agency microdata and making research data (without PII) available for analysis by authorized researchers and agency staff. It also includes a web-based metadata repository for authorized users. The ADRF environment is compliant with the Federal Risk and Authorization Management Program (FedRAMP), the government-wide program that provides a standardized approach to security assessment, authorization, and monitoring of cloud-based products and services.[11]

The ADRF prototype has been constructed and is being tested by participants in the Applied Data Analytics training program offered jointly by the University of Chicago, New York University, and the University of Maryland.[12] This innovative program has been developed to give working professionals the opportunity to develop key computer science and data science skills to enable data-driven decisions in public policy and the public sector. The first class launched in February 2017 with a cohort of students comprised predominantly of staff from state agencies. Through the course, students submitted real agency data to the ADRF environment and used that data to answer contemporary policy questions of interest to their agency.

Key to the success of a cloud-based data platform is that all stakeholders see the environment as a legitimate enterprise that is of value to their organization with imposing undue risks or costs. This includes the commitment of public data providers to sharing data and integrating findings into decision making. Issues of data privacy and security are of primary importance to all involved and efforts should be made to adhere to current and evolving best practices. Other requisite factors are organizational management that addresses political and financial sustainability, efficient legal compliance with data sharing and other agreements, and ensuring appropriate scientific rigor and dissemination of research products and results.

Transferring local data with PII to the Census Bureau

Typically, data providers submit to the Census Bureau several pieces of direct personally identifiable information (PII) for the assignment of Protected Identification Keys (PIK) within the Data Linkage Infrastructure. However, some data providers may be unable to transfer PII to the Census Bureau. This may be the case when a data provider has given masked IDs or already-linked data to a principal investigator. In one such case, the Census Bureau is working to replicate the masking, or "hashing," process and apply to other relevant datasets from within the Census Bureau infrastructure. Additionally,

[11] FedRAMP: https://www.fedramp.gov/
[12] Training Program in Applied Data Analytics: http://www.applieddataanalytics.org/

the Census Bureau is interested in exploring secure multiparty computation methods as a way to process data without transferring data to its infrastructure.

Review by data providers

While the Census Bureau conducts a disclosure review of all output to prevent the identification of individuals or businesses, as well as a review to ensure output falls within the approved proposal scope, sometimes data providers also require review of products using their data, such as journal articles and working papers. The level of this review varies and might include data owners being informed of publications, data owners receiving advance copies of publications, data owners having the opportunity to comment on products prior to publication, or data owners reviewing and approving product content. Some datasets are already restricted to a single project with strong input from the data owner. In other cases, data owners have requested additional time for review, additional analysis into unanticipated findings, or asked for more context around certain statistics. In such cases, the Census Bureau has worked with data providers without censoring content.

Summary of Challenges

The federal- and local-level challenges addressed here are not unique to this project. Most issues have been raised by others. Efforts have been made to address these challenges and more such efforts are currently underway, but much work remains to achieve sufficient capacity to expand data analysis for evidence-based policy making. From the specific challenges highlighted here, several dominant themes emerge. First, current staffing and other issues of capacity place constraints on advancing the work. The development and maintenance of sufficient metadata would require initial and ongoing investments by the Census Bureau and the data providers, which are comprised of agencies at the federal, state, and local level, as well as private companies. Second, the time and resultant cost of executing the numerous legal and data sharing agreements is burdensome for all parties. Exploring avenues for efficiency gains in the process through templates, model legislation and agreements, and training of legal and agency staff could be of benefit. Also central to this process is ensuring that data providers perceive a mutual benefit from these projects and have a high level of trust in the process and data environment. Third, requiring physical presence of researchers in specific facilities imposes additional time and costs and also reduces accessibility for some otherwise qualified researchers. Certain data sources may always justify the precautions of on-site data analysis, but options that take advantage of current technology with cloud-based solutions may be feasible in many circumstance. While less central to advancing this work, additional issues (such as introducing greater transparency processes for establishing qualified researchers, testing new methods for transmitting data without conventional PII to the Census Bureau, and protocols for review of research products) are also worthy of attention.

Conclusion

The purpose of this report is to describe demand by researchers from a range of institutional auspices for the data sources held by the Census Bureau and to highlight the potential for advances in evidence-based policy making generated by research projects utilizing these data. To do this, we drew primarily from the 45 project proposals submitted in response to the "Using Linked Data to Advance Evidence-Based Policymaking: Helping Projects Utilize the U.S. Census Bureau Linkage Infrastructure" RFP released in August 2016 by Chapin Hall at the University of Chicago in cooperation with the U.S. Census Bureau.

Demand for combining local data with federally held administrative data sources originates from governmental and research communities, including local, state, and federal entities; academic and research institutions; and in collaborations with nonprofit organizations. There is interest in utilizing data held at the Census Bureau to inform policy decision making across all levels of government: school districts, cities, counties, states, regional coalitions, and federal agencies. In project proposals, researchers addressed issues of education, employment, and earnings, early care and education, public assistance programs, housing and homelessness, health, criminal justice, immigrant and refugee populations, child custody and child support, public safety, taxation, and regulation. In order to study these topics researchers sought to use multiple sources of data held at the Census Bureau: taxes, wages and employment, Census Bureau surveys, public assistance programs (TANF, Medicaid/Medicare, SNAP, housing), residential mobility, and mortality. The breadth of proposed study topics and data sources reveals that many crucial policy issues stand to gain from an increased capacity to match local and federally held data for research purposes within a secure, privacy-protecting environment. The potential to assess long-term outcomes of public program participants and responses to policy changes is particularly promising.

Key Challenges and Learnings

From the RFP process, six pilot projects were selected for immediate implementation, which began in November 2016. The experience of implementing the pilots has identified several challenges to date that are common to most, if not all, of the pilot projects. First, it is essential have a good understanding during the proposal development process of the data sources held at the Census Bureau. Currently, in-depth information about the datasets can only be obtained through personal contacts with staff at the Census Bureau, a model that is not scalable for an expansion of this type of research. Second, the process by

which individual researchers obtain access to Census Bureau data resources (obtaining special sworn status) is lengthy, sometimes opaque, and can contribute to project delays. Third, getting the necessary data sharing agreements and permissions in place is a lengthy process that involves multiple parties for each project. This process requires a significant time investment from all parties and necessitates a strong collaborative partnership with data providers to achieve success. Finally, the costs for researcher access and data matching work at the FSRDCs are not standardized and vary dramatically by FSRDC site. The cost of executing a project is somewhat determined by the geographic location. Further, depending on their location, some researchers would have to undertake significant travel in order to access an FSRDC.

Recommendations

To address the key challenges outlined above, we recommend the following:

- Invest in increased capacity at the federal level to support projects combining local data with federal and federally held data sources. Specific areas for growth include: developing and maintaining sufficient metadata to support quality proposal development, develop clear and consistent processes for assessing researcher qualifications and obtaining necessary permissions, and ensuring sufficient staff to accommodate the additional work flow.
- Investigate and develop strategies for navigating legal and data sharing issues, streamlining processes whenever possible while also recognizing the importance and need for developing strong and collaborative relationships between all partners and data providers.
- Explore options for cloud-based data administrative data research facilities that can increase accessibility while maintaining a high level of privacy and security.
- Plan for sustainability to support the growth of efforts to combine data across sources for the evaluation of public programs and to advance policymaking. This could include exploring options like fee structures and cost-sharing models; having alternate hosts for select sources of federally held data; more direct participation of data providers, such as states, in the Census Bureau Data Linkage Infrastructure; or an administrative data research facility that incorporates federal, state, local and other public and private data sources.

Matched administrative data systems greatly increase our ability to generate actionable information about policies at low cost, because they provide richer information on outcomes across domains and give researchers and government officials a single access point to relevant data held by others. However, the full potential of these systems has not yet been realized. Combining evaluation data of interest with other datasets held at the Census Bureau can dramatically increase evidence-building capacity—if the Census Bureau's Data Linkage system is sufficiently resourced and expanded to support more high-quality evaluation and programmatic research.

Appendix A. Submitted Project Proposals

This table lists projects submitted as full proposals and letters of interest in response to the "Using Linked Data to Advance Evidence-Based Policymaking: Helping Projects Utilize the U.S. Census Bureau Linkage Infrastructure" RFP. Author names, affiliations, and locality were withheld at author request. State/Region indicates the state or region of the proposal author's organization. This often corresponded with the geographic region of the "local" data in the project but this was not always the case; some researchers proposed national analyses, utilize samples from specific research studies, or represent research-government partnerships that cross state lines.

Title	Author(s)	Organization	State/Region
A Tool for Head Start and Related Programs to Assess Community Needs	Withheld	National Head Start Association	Virginia
Adult Labor Market Outcomes of Chicago Public School Students: Pathways from K-12 to Work	Derek Neal	University of Chicago	Illinois
Analysis of the Impact of Behavioral and Physical Healthcare on Risk of Suicide Death among Medicaid and Medicare Enrollees	Withheld	Withheld	Withheld
Does Eviction Cause Poverty? Quasi-experimental Evidence from Cook County, IL	Daniel Tannenbaum Winnie van Dijk	University of Nebraska University of Chicago	Nebraska Illinois
Effects of the Child-Parent Center Preschool-to-3rd Grade Program on Economic Well-Being and Health Status in Two Longitudinal Cohorts	Arthur Reynolds Suh-Ruu Ou	University of Minnesota-Twin Cities	Minnesota
Experimental Estimates of the Long-run Impacts of Welfare Reform on Participants and their Children	Jordan Matsudaira Pauline Leung Zhuan Pei	Cornell University	New York
Food Insecurity in Michigan: Using Cluster Analysis and Geographic Information Systems to Identify Community Needs	Stephen Borders Kait Skwir	Grand Valley State University Food Bank Council of Michigan	Michigan
Health at Birth and Later Life Outcomes: Evaluating the Returns to Policy-driven Early Health Investments	Laura Wherry Sarah Miller	University of California, Los Angeles University of Michigan	California Michigan

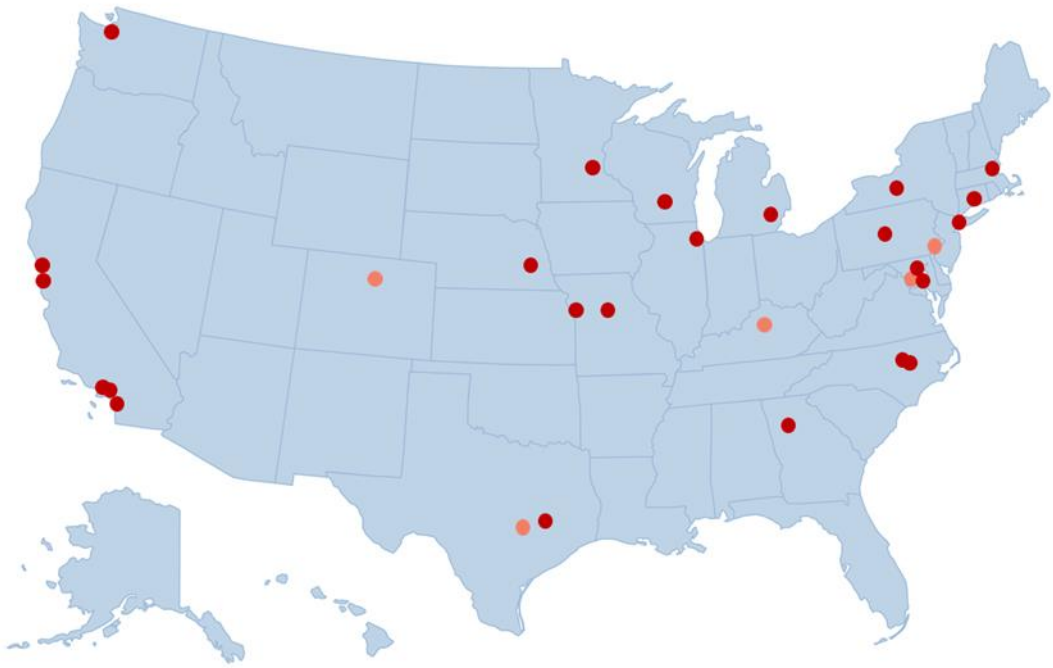
Title	Author(s)	Organization	State/Region
High School, College and Career Pathways: Impact on Economic Outcomes in Adulthood	Rachel Durham	Baltimore Education Research Consortium at Johns Hopkins University	Maryland
How Health Habits Influenced the Level of Income Inequalities in the US Counties: Longitudinal Analytics (2000-2015)	Hossein Zare	University of Maryland University College Johns Hopkins University	Maryland
Identifying and Assessing New Mexico's Multi-System Families	Jennifer Ramo Lisa Alejandro	New Mexico Appleseed	New Mexico
Impact of Mobility on Community Partnerships for Children	Mary Klos	Achieve Brown County	Wisconsin
Impact of the Early College High School Model on Lifetime Outcomes	Julie Edmunds Fatih Unlu	University of North Carolina at Greensboro RAND Corporation	North Carolina
Interrupting Intergenerational Poverty in Detroit	Jeannine La Prad	Corporation for a Skilled Workforce	Michigan
Linking ACS and NAEP to Examine Educational Achievement in Appalachia at Grades 4, 8, and 12	Withheld	Withheld	Withheld
Linking the PIC Dataset and Local School Data to Measure Community Learning Center Impact	Kara Byrne Ivis Garcia Zambrana	University of Utah	Utah
Long-term Income, Well-being and Public Cost Outcomes of Public Assistance Recipients	Daniel Flaming Halil Toros	Economic Roundtable	California
Marginal Tax Rates and Employment Among Low-income Populations	Angela Rachidi	American Enterprise Institute	Washington DC
Measuring the Impact of a Near-universal Child Support System on Payment of Child Support	Daniel Schroeder	Ray Marshall Center, University of Texas at Austin	Texas
National Longitudinal Transition Study 2012	Michael Bryan Yumiko Sekino	RTI International US Department of Education	Washington DC
Neighborhoods, Assets, and Credit: Another Look at the Moving-to-Opportunity Project	Erik Hembre	University of Illinois at Chicago	Illinois
Outcomes under the Post-9/11 GI Bill	Mark Schneider Carrie Wofford	American Institutes for Research Veterans Education Success	Washington DC
Policies to Improve Health Equity within US States	Eoghan Brady David Bishai	Johns Hopkins School of Public Health	Maryland
Program Utilization by Formerly Criminalized Youth: Linking Juvenile Recidivism Data to Data Held by US Census Bureau	Andrew Clark Ashley Provencher	Institute for Municipal and Regional Policy, Central Connecticut State University	Connecticut

Title	Author(s)	Organization	State/Region
Proposal to Study Public Safety Calls for Service to Improve Disaster Response and Preparedness Capabilities	Withheld	Withheld	Maryland
Quad Cities Data Warehouse Economic Achievement Gap Research	Alex Kolker	United Way of the Quad Cities Area	Illinois, Iowa
RegData: Quantifying Regulation and the Regulatory Process	Patrick McLaughlin Sarah Jenslawski	Mercatus Center at George Mason University	Virginia
Socioeconomic Disparities in Parent Enforcement of the IDEA	Rebecca Johnson	Princeton University	New Jersey
Supporting Veterans to Enter the Workforce	Withheld	Withheld	Withheld
State-level Variations in CCDF Policy: What Makes a Difference for Children and Families?	Jade Jenkins	University of California, Irvine	California
Testing Predictive Approaches to Policymaking in a Large Urban County	Withheld	Withheld	Withheld
The Causal Impact of Naturalization on the Economic Integration of Immigrants	Jens Hainmueller Duncan Lawrence	Stanford University Immigration Policy Lab	California
The Effect of Custodial Arrangements on Parent and Child Outcomes	Daniel Tannenbaum Manasi Deshpande	University of Nebraska University of Chicago	Nebraska Illinois
The Intergenerational Consequences of Incarceration	Matthew Pecenco Samuel Norris	University of California at Berkeley Northwestern University	California Illinois
The Long-Run Impacts of Financial Aid	Withheld	Withheld	Withheld
The Minimum Wage Study (MWS) at the University of Washington	Scott Allard	University of Washington	Washington
The Persistent Effect of Childhood Lead Exposure on Student Performance and Labor Market Outcomes	Marie Lynn Miranda John Killeen	Children's Environmental Health Initiative, Rice University Data Works NC	North Carolina
Understanding SNAP Recipients and Unenrolled SNAP Eligibles in Philadelphia using Linked Data	Candice Dias Carolyn Brown	Pennsylvania Department of Human Services City of Philadelphia	Pennsylvania
Understanding the Viability of the Federal Poverty Level as an Eligibility Screen for Head Start and Other Federal Poverty Programs	Withheld	National Head Start Association	Virginia

Title	Author(s)	Organization	State/Region
Unlocking Refugee Potential: The Impact of Policies and Programs on Refugee Economic Integration in the United States	Jeremy Weinstein Andrea Dillon	Stanford University	California
Using Linked Data to Advance Employment Studies	Richard Hendra	MDRC	New York
Using Linked Data to Examine the Trajectories and Service Utilization of Families and Children Experiencing Homelessness	Andrew Reynolds Vikki Cherwon	University of North Carolina at Charlotte	North Carolina
Using Linked Data to Investigate Improvements in Predictive Modeling Accuracy	Withheld	Withheld	Withheld
Using Multiple Discontinuities to Estimate Broad Effects of Public Need-based Aid for College	Drew Anderson	Wisconsin HOPE Lab, University of Wisconsin-Madison	Wisconsin
When White Schools Disappear	Habiba Ibrahim Odis Johnson Jr	Washington University in St. Louis	Missouri

Appendix B. Location of Federal Statistical Research Data Centers

Location of Federal Statistical Research Data Centers



Source: U.S. Census Bureau, retrieved from https://www.census.gov/fsrdc

Federal Statistical Research Data Centers (RDC) are currently operating in these locations:

1. Atlanta, GA (Atlanta RDC)
2. Cambridge, MA (Boston RDC)
3. Berkeley, CA (California RDC at Berkley)
4. Irvine, CA (California RDC at Irvine)
5. Stanford, CA (Stanford RDC)
6. Los Angeles, CA (California RDC at UCLA)

7. Los Angeles, CA (California-USC RDC)
8. Suitland, MD (Census Bureau Head Quarters RDC)
9. Lincoln, NE (Central Plains RDC)
10. Chicago, IL (Chicago RDC)
11. Kansas City, MO (Kansas City RDC)
12. Lexington, KY (Kentucky RDC) *
13. College Park, MD (Maryland RDC)
14. Ann Arbor, MI (Michigan RDC)
15. Minneapolis, MN (Minnesota RDC)
16. Columbia, MO (Missouri RDC)
17. New York, NY (NYRDC-Baruch)
18. Ithaca, NY (NYRDC-Cornell)
19. Seattle, WA (Northwest RDC)
20. State College, PA (Penn State RDC)
21. Philadelphia, PA (Philadelphia RDC) *
22. Boulder, CO (Rocky Mountain RDC) *
23. College Station, TX (Texas RDC)
24. Durham, NC (Triangle RDC - Duke)
25. Research Triangle Park (Triangle RDC – RTI)
26. Austin, TX (Texas RDC – UT Austin) *
27. Washington, DC (Georgetown RDC) *
28. Madison, WI (WiscRDC)
29. New Haven, CT (Yale RDC)
30. Urbana Champaign, IL (UIUC FSRDC) *

* Projected to open in 2017

About Chapin Hall

Chapin Hall is an independent policy research center at the University of Chicago focused on providing public and private decision-makers with rigorous data analysis and achievable solutions to support them in improving the lives of society's most vulnerable children. Chapin Hall partners with policymakers, practitioners, and philanthropists at the forefront of research and policy development by applying a unique blend of scientific research, real world experience, and policy expertise to construct actionable information, practical tools, and, ultimately, positive change for children, youth, and families.

Established in 1985, Chapin Hall's areas of research include child and adolescent development; child maltreatment prevention; child welfare systems; community change; economic supports for families; home visiting and early childhood initiatives; runaway and unaccompanied homeless youth; schools, school systems, and out-of-school time; and youth crime and justice.

1313 East 60th Street
Chicago, IL 60637

773-256-5100
www.chapinhall.org

Chapin Hall at the University of Chicago
Policy research that benefits children, families, and their communities