Accessing Confidential Economics, Demographic, and Health Data Through the California Census Research Data Center (RDC)

Till von Wachter, Director, RDC, tvwachter@econ.ucla.edu

John Sullivan, Administrator, RDC, john.sullivan@census.gov

Table of Contents

Table of Contents

Introduction: How to Access Data at RDC?

- RDC: secure data lab to access confidential government data
- Rich Amount of Data:

  1. Demographic (individual) data (Census) 2. Business data (Census)

  3. Health data (NCHS) 4. Labor data (BLS) 5. Economic accounts data (BEA) 6. Administrative Data (e.g., earnings)
- How to use RDC:

  project proposal; agency approval; research in lab; disclosure process
- 3 Goals of Talk:

  1) Make you aware of data; 2) Tell you how to access data; 3) Financing

Some Key Takeaways on RDC

1. Some Key Points Regarding Data

- There are some low hanging fruits (Demographic; Business, Health)
- Health data in particular is a big untapped resource

2. RDC Network and Available Data Likely to Grow

- Increasing number of data sets. Increasing number of branches and projects.

3. Getting project approval differs by data source and agency

- Good to plan ahead and make an investment into data access
- We are there to help – hope to do even more in future

4. Funding situation at UCLA

- Graduates students get in for free if their unit participates. Others have to pay.
- *UCLA last RDC that is mainly fee based. We are taking steps to change that.*

Background on RDC Network

Census Bureau administers a network of RDCs Across U.S.

- There are currently 30 RDCs (and branches)
- Large number of active research projects (200+)

An Increasing Number of Agencies is Using RDC Network

- Census is joined by NCHS, AHRQ, BLS, BEA
- Changed name to *Federal Statistical Research Data Centers (FSRDC)*

Goal of RDC is to make data accessible while safeguarding confidentiality

- Stringent rules of access and disclosure *(more on this below)*

Rules Governing Data Access Differs Across Agencies

- The law allows Census to give access to data to improve Census data products. Access is simpler for NCHS and AHRQ data.

Part 1: Data Available in the RDC

Overview of Types of Data Sources

1. **Demographic Data (Census Bureau)**
   - Ex: Decennial Census, ACS, CPS, NLMS, etc.
2. **Merged and Administrative Data Sets**
   - Ex: Matched Decennial Censuses, Survey Data to Earning Histories, Longitudinal Employer Household Dynamics (LEHD)
3. **Economic Data (Census Bureau)**
   - Ex: Economic Census, Annual Survey of Manufacturing, Longitudinal Business Database (LBD), etc.

Overview of Types of Data Sources, cont.

4. Health Data (National Center for Health Statistics, Agency for Healthcare Research and Quality)

- Ex: finer geographic detail; finer detail on race/ind/occ; added information

5. Labor Data (Bureau of Labor Statistics)

- Ex: NLSY with geocodes; occupation injury statistics

6. Economic Accounts Data (Bureau of Economic Advisors)

- Ex. Annual Survey of U.S. Direct Investment Abroad

1. Demographic Data – Why Use the RDC?

- Micro-data with detailed geography
  - Tract level for most (address, latitude/longitude for some)
  - Not available in public micro-data
- Less severe top coding
- Some datasets have additional variables
- Opportunities for individual level linkages (PIKs)
- Potential for "unswapped" data
- Not suitable venue for a "special tabulation"

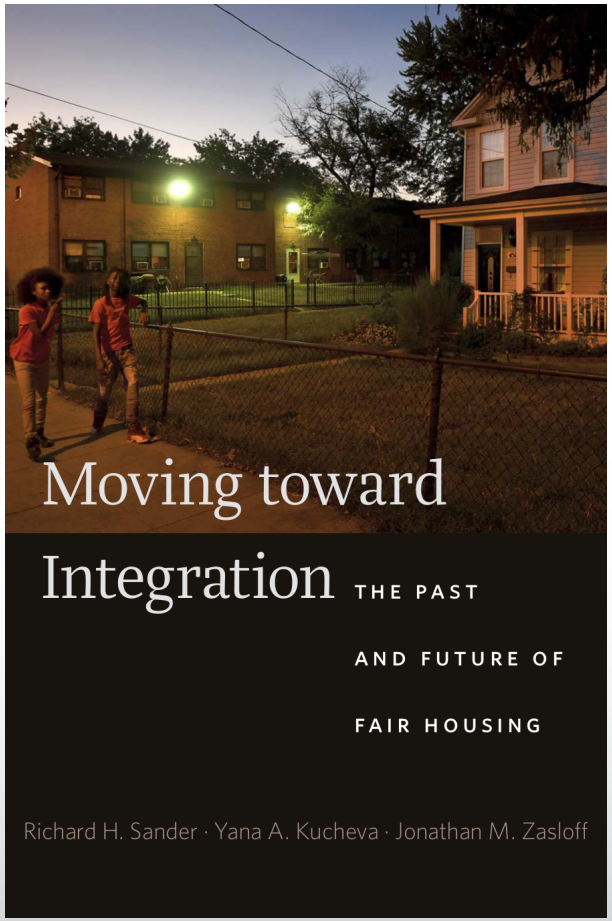
1. Demographic - Available Data (1)

- Decennial Surveys
  - 1940 - 2010
- American Community Survey (ACS)
  - Annual microdata, 1996-2019
- Current Population Survey (CPS)
  - Various Supplements (including March ASEC)

1. Demographic - Available Data (2)

- Survey of Income and Program Participation (SIPP)
- American Housing Survey (AHS)
- National Crime Victimization Survey (NCVS)
- National Longitudinal Mortality Study (NLMS)
- National Longitudinal Surveys (NLS)
  - Young/Mature Men/Women

Example: Historical Segregation at Neighborhood Level



- Sander, Kucheva and Zasloff.
- Decennial Census 1960-2010
  - Block-level location of households by race, decennial long-form
  - Components of historical segregation and neighborhood change

2. Administrative and Matched - Available Data (1)

- CPS and SIPP extracts matched to SSA's earnings records
  - SER, DER, MBR
- HUD - Moving to Opportunity
- WIC/SNAP
- Census Numident
- UMETRICS
  - Information on awards, wage payments from awards to university research employees, vendor purchases and the unit performing the funded research for 26 universities.
  - Linked to internal Census Bureau data products

2. Administrative and Matched - Available Data (2)

- PIKs (Protected Identification Keys)
- Linked Decennial Censuses (1940, 2000, 2010) + ACS (2006-2019)
- MAF-ARF
  - Annual address level location records for large portion of population
- MAFX
  - Master Address File – cumulative universe of US addresses (some unit characteristics)
- Link external data to restricted data on the individual level

2. Administrative & Matched – Integrated Worker-Firm Data

- Longitudinal Employer Household Data (LEHD)
  - Administrative quarterly employment and earnings records for workers and firms from state's Unemployment Insurance systems merged with demographic data from SSA
  - Requires Census, IRS, and SSA approval

Example: Long-Term Effects of Local Policies

War on Poverty Project

- Bailey, et. al.
- Decennial Census, ACS and Numident
  - Numident records of place of birth for massive sample matched to demographic data.
  - Participation in Head Start had positive effects on economic self-sufficiency, schooling and college completion.

3. Business Data– Why Use the RDC?

- Establishment-level data
  - Essentially no publically available micro-data
- Detailed geography information
- Establishment – firm linkages possible
- Longitudinal linkages possible
- Linkage across economic and mixed data (e.g., worker-level data) products

3. Business - Available Data (1)

- Economic Census
  - Annual Survey of Manufactures, Annual Survey of Retail Trade, Annual Survey of Services, Monthly Wholesale Trade Survey.
- Quarterly Financial Report (QFR)
- Survey of Business Owners (SBO)
- Medical Expenditure Panel Survey Insurance Component (MEPS-IC) → AHRQ

3. Business - Available Data (2)

- Longitudinal Business Database and ILBD (integrated LBD)
  - Basic information on the universe of establishments
    - Links to parent firms
    - Birth/death dates, longitudinal links
    - Measures of size, revenue (for SUs), industry, LFO
    - Updated annually
- Business Register (Standard Statistical Establishment Listing)
  - Aids linkage to other economic data products, some additional information
  - Can accommodate linkage to external data (i.e. Compustat, records with business name/location)

3. Business - Available Data (3)

- Annual Capital Expenditures Survey (ACES)
- Business Research & Development and Innovation Survey (BRDIS)
- Manufacturers' Shipments, Inventories, and Orders (M3)
- Survey of Pollution Abatement Costs and Expenditures (PACE) and Manufacturing Energy Consumption Survey (MECS)

3. Business - Available Data (4)

- Commodity Flow Survey (CFS)
- Longitudinal Foreign Trade Transactions Database (LFTTD)
- Kauffman Firm Survey
- Longitudinal Employer Household Data (LEHD)
  - Administrative quarterly employment and earnings records for workers and firms from state's Unemployment Insurance systems merged with demographic data from SSA
  - Requires Census, IRS, and SSA approval

Example: Research from the UCLA RDC

STRATEGIC MANAGEMENT JOURNAL

SMS Strategic Management Society

RESEARCH ARTICLE

Human capital, parent size, and the destination industry of spinouts

Mariko Sakakibara ✉, Natarajan Balasubramanian

First published: 07 November 2019 | https://doi.org/10.1002/smj.3108

Funding information Whitman School of Management at Syracuse University; Academic Senate of the University of California, Los Angeles; Harold Price Center for Entrepreneurial Studies at UCLA Anderson School of Management; Kauffman Foundation; Alfred P. Sloan Foundation; National Institute on Aging Grant, Grant/Award Number: AG018854; National Science Foundation Grants, Grant/Award Numbers: ITR-0427889, SES-0339191, SES-9978093

- Sakakibara and Balasubramanian
- LEHD linked to LBD
  - Tracks survival and performance of spinout firms as function of founder characteristics and parent firm.

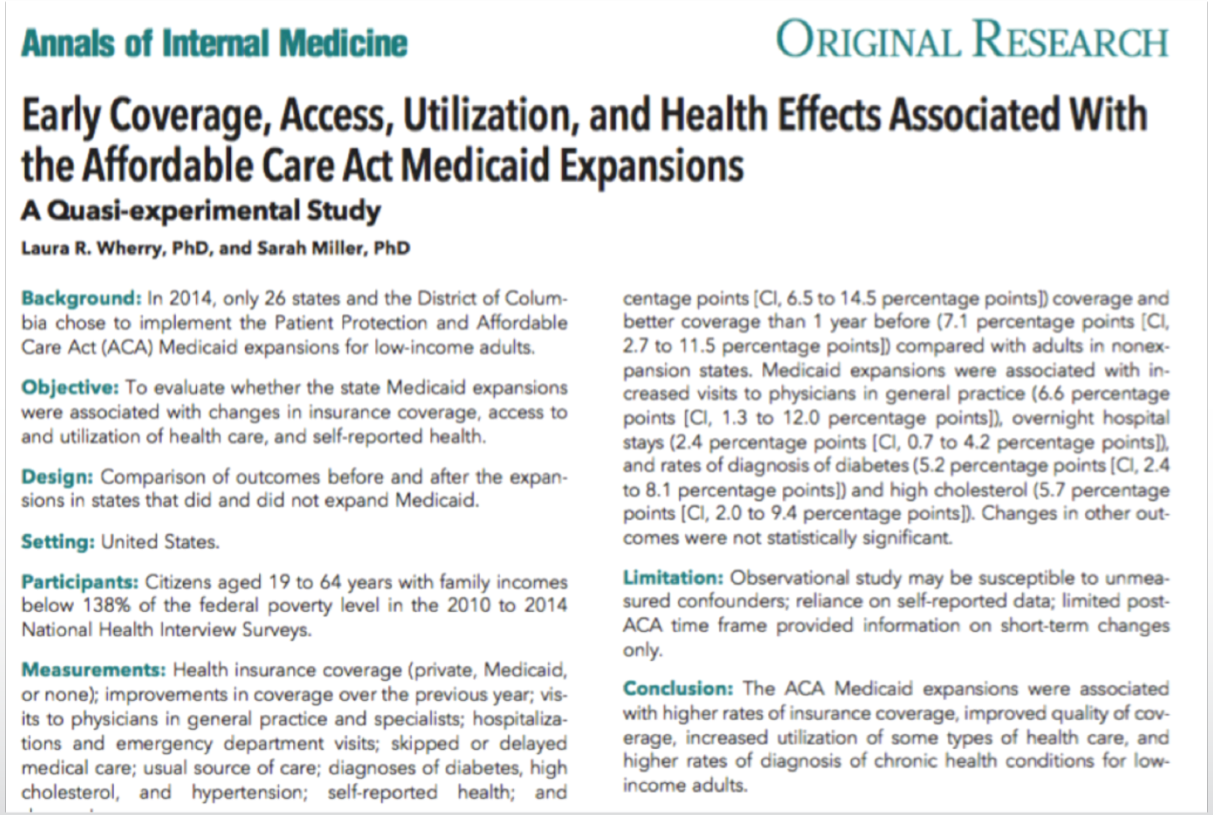
4. Health Data - Why use the RDC?

- More detailed level of geographical information
- Data linkages available for NCHS surveys
  - Mortality
  - Medicare meta-data
  - Social Security Benefits
- Greater detail available for key variables
  - Race
  - Disease codes
  - Industry and occupation codes

4. Health Data – Available Data

- AHRQ (Agency for Healthcare Research and Quality)
  - Medical Expenditures Panel Survey (MEPS)
- NCHS (National Center for Health Statistics)
  - National Health Interview Survey (NHIS)
  - National Health and Nutrition Examination Survey (NHANES)
  - National Survey of Family Growth (NSFG)
  - National Vital Statistics System (NVSS)
  - National Health Care Surveys – NAMCS, NHAMCS, NHDS, NNHS, NNAS, NSRCF, NSLTCP

Example: Research from the UCLA RDC



Annals of Internal Medicine ORIGINAL RESEARCH

Early Coverage, Access, Utilization, and Health Effects Associated With the Affordable Care Act Medicaid Expansions

A Quasi-experimental Study

Laura R. Wherry, PhD, and Sarah Miller, PhD

Background: In 2014, only 26 states and the District of Columbia chose to implement the Patient Protection and Affordable Care Act (ACA) Medicaid expansions for low-income adults.

Objective: To evaluate whether the state Medicaid expansions were associated with changes in insurance coverage, access to and utilization of health care, and self-reported health.

Design: Comparison of outcomes before and after the expansions in states that did and did not expand Medicaid.

Setting: United States.

Participants: Citizens aged 19 to 64 years with family incomes below 138% of the federal poverty level in the 2010 to 2014 National Health Interview Surveys.

Measurements: Health insurance coverage (private, Medicaid, or none); improvements in coverage over the previous year; visits to physicians in general practice and specialists; hospitalizations and emergency department visits; skipped or delayed medical care; usual source of care; diagnoses of diabetes, high cholesterol, and hypertension; self-reported health; and

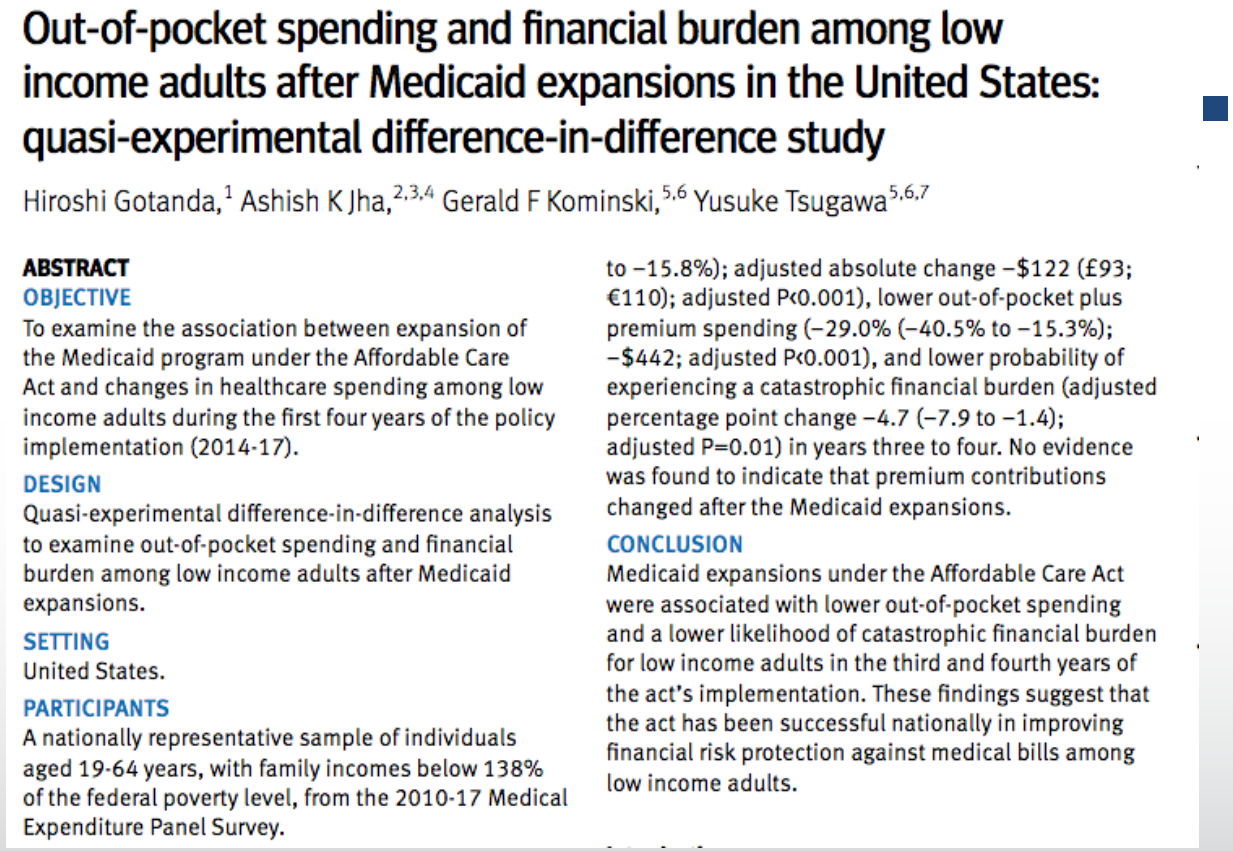
centage points [CI, 6.5 to 14.5 percentage points]) coverage and better coverage than 1 year before (7.1 percentage points [CI, 2.7 to 11.5 percentage points]) compared with adults in nonexpansion states. Medicaid expansions were associated with increased visits to physicians in general practice (6.6 percentage points [CI, 1.3 to 12.0 percentage points]), overnight hospital stays (2.4 percentage points [CI, 0.7 to 4.2 percentage points]), and rates of diagnosis of diabetes (5.2 percentage points [CI, 2.4 to 8.1 percentage points]) and high cholesterol (5.7 percentage points [CI, 2.0 to 9.4 percentage points]). Changes in other outcomes were not statistically significant.

Limitation: Observational study may be susceptible to unmeasured confounders; reliance on self-reported data; limited post-ACA time frame provided information on short-term changes only.

Conclusion: The ACA Medicaid expansions were associated with higher rates of insurance coverage, improved quality of coverage, increased utilization of some types of health care, and higher rates of diagnosis of chronic health conditions for low-income adults.

- Laura Wherry and Sarah Miller
  - National Health Interview Survey with restricted state identifiers
  - Effect of ACA state Medicaid expansions on insurance coverage, access and utilization of health care and self-reported health

Example: Research from the UCLA RDC



Out-of-pocket spending and financial burden among low income adults after Medicaid expansions in the United States: quasi-experimental difference-in-difference study

Hiroshi Gotanda,[1] Ashish K Jha,[2,3,4] Gerald F Kominski,[5,6] Yusuke Tsugawa[5,6,7]

ABSTRACT

OBJECTIVE

To examine the association between expansion of the Medicaid program under the Affordable Care Act and changes in healthcare spending among low income adults during the first four years of the policy implementation (2014-17).

DESIGN

Quasi-experimental difference-in-difference analysis to examine out-of-pocket spending and financial burden among low income adults after Medicaid expansions.

SETTING

United States.

PARTICIPANTS

A nationally representative sample of individuals aged 19-64 years, with family incomes below 138% of the federal poverty level, from the 2010-17 Medical Expenditure Panel Survey.

to −15.8%); adjusted absolute change −$122 (£93; €110); adjusted P<0.001), lower out-of-pocket plus premium spending (−29.0% (−40.5% to −15.3%); −$442; adjusted P<0.001), and lower probability of experiencing a catastrophic financial burden (adjusted percentage point change −4.7 (−7.9 to −1.4); adjusted P=0.01) in years three to four. No evidence was found to indicate that premium contributions changed after the Medicaid expansions.

CONCLUSION

Medicaid expansions under the Affordable Care Act were associated with lower out-of-pocket spending and a lower likelihood of catastrophic financial burden for low income adults in the third and fourth years of the act's implementation. These findings suggest that the act has been successful nationally in improving financial risk protection against medical bills among low income adults.

- Gotanda, et. al.
  - MEPS-HC
  - Medicare expansion under ACA lowered cost burdens and risk of catastrophic financial burden for low income adults.

5. New Data from Bureau of Labor Statistics

- National Longitudinal Surveys of Youth (NLYS79 and NLSY97)
  - NLSY with regional identifiers
- Survey of Occupational Injuries and Illnesses
- Application process structured similar to NCHS
  - Contact and apply through BLS, but access data through RDC

6. Data from Bureau of Economic Analysis

- Data
  - Annual Survey of U.S. Direct Investment Abroad
  - Annual Survey of Foreign Direct Investment in US
  - Survey of Transaction in Selected Services and Intellectual Property
- Proposals reviewed by BEA
- Potential for firm level linkage to census economic data

Ongoing Developments

- Virtual RDC (VRDC)
  - Census Bureau has begun a remote access pilot allowing researchers access their RDC project from a location other than the physical RDC.
  - Pilot widely expanded during COVID closure, expected to continue.
  - No T26 (FTI) and some limits on NCHS-provided data.
- UCLA FSRDC Move
  - Moving to newly constructed office in the Social Science Computing hub (Luskin Public Affairs Building 2nd floor)

Statistics of Income

- SOI Laptops for IRS research in the FSRDCs
  - IRS approved projects
  - FSRDCs as location of research
  - Presumably universe of tax records
    - 1040 Individual Tax Returns, 1099 Information Returns Master File
- Public Call for Proposals
  - https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program

Part 2: How To Access RDC Data

Application Process: Census vs. NCHS, AHRQ, BLS, BEA

- Census: proposals are submitted to the RDC administrator
  - Census reviews proposals for both content & benefits from research
- NCHS/AHRQ/BLS/BEA: proposals submitted directly to agency – does not go through the RDC administrator
  - Must still contact RDC administrator before
  - Generally, easier to apply and applications are processed more quickly than projects using Census, IRS or other agency data
  - Fees paid to NCHS for data extracts

Ease of Access Can Vary Between Data Sets:

“Cookie cutter projects"

- Demographic data (only Census approves)
- Business data (Census & IRS approve)
- Health data (only NCHS or AHRQ approves, but no scientific merit review)
- Easy to add in public data sources (as long as specified in advance)

Higher hanging fruits

- Merge between various data sets
- Merge outside confidential data
- Merge data from various agencies (Ex: LEHD)

Overview: Two Approaches to Data Access

Basic Procedure for Census Data Proposal (Demographic and Business Data):

1. Get idea & check data; talk to RDC administrator & director
2. Write proposal explaining research idea, statistical analysis, and data
3. Come up with two statistical "Benefits for the Census Bureau"
4. Submit proposal for review with the RDC administrator, make revisions
5. Agency reviews proposal, may ask for revisions
6. In the meantime, fill out paperwork for Special Sworn Status (SSS)
7. Once project is approved, work in RDC. Output obtained in disclosure review

Key Differences in Format of Proposal for NCHS, AHRQ, BLS, BEA:

1. Proposal does not require benefits, but requires a specific list of variables.
2. Only formal review, no content review. Review times much faster.

Discuss Some Important Practical Issues

1. Practical considerations for writing proposals

2. What constitutes a “Benefit for the Census Bureau”

3. Questions About Special Sworn Status

4. Information About Confidentiality and the Disclosure Process

Suggestions for Census Data Proposals

1. Contact the RDC Administrator with idea
2. Get started on proposal sooner rather than later
   - Benefits often emerge as proposal is developed
   - Written for a data expert rather than a content expert
   - Description limit is 15 pages single spaced (30 pages double)
3. Plan ahead:
   - 3-6 months (health) to a 6-12 months (demog. & business)

Proposal Outline

- Intro (1-3 pgs.)
  - Overview of benefits; describe research question; brief lit. rev.; overview of research plan and data
- Methodology (8-9 pgs.)
  - Detailed model specification, key variables, how data will be used in estimation; methods to complete benefits
- Data (1-3 pgs.)
  - Bureau-provided data; External Data
  - Linkage
- Output and Disclosure Risk (1-3 pgs.)
  - Model-based output (emphasized)
  - Tabular output
  - Technical memos
  - Disclosure risk and mitigation
- Duration and Funding (<1 pg.)

Predominant Purpose - Benefits

- The predominant purpose of projects approved under Title 13 is to provide benefits to the Census Bureau.
- 13 benefit criteria (IRS only recognizes 9 of the 13).
- #11 - Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;
  - All projects claim #11 and usually only one additional benefit.
  - E.g. estimating non-response; develop weighting strategy; improve imputation; understand/improve data quality; construct/verify/improve sampling frames; evaluate concepts and practices of data collection

Benefits - Examples

- Economic – Project uses ASM, CMF and LBD to study firm exit and capital misallocation
  - Benefit 1: *Understanding and/or improving data quality*
    - Utilizes edit/impute flags to examine unit and item non-response in the ASM/ CMF separately for surviving and near death firms.
  - Benefit 2: *Enhancing the data collected*
    - Develops a model to impute select missing fields. Imputation model is novel for its consideration of information on subsequent firm exit.

Benefits - Examples

- Demographic – Project uses Decennial Census and ACS to study racial residential segregation
  - Benefit 1: *Increasing the utility of data for analyzing public policy and/or demographic, economic or social conditions*
    - Demonstrates the importance of the residential mobility questions on the Decennial and ACS for understanding patterns of racial segregation.
    - Documents comparability issues in survey items and geographic boundaries.
  - Benefit 2: *Preparing estimates of population*
    - Develops models of racial change in census tracts and metropolitan areas.

Notes on IRS Review

- IRS reviews all proposals for Federal Tax Information (FTI)
  - Most economic data contain FTI (LBD, EC, LFTTD, etc.)
  - A small amount does not (i.e. raw IMP/EXP)
  - There are Title 26 (has FTI) and non-Title 26 versions of the LEHD ICF
    - T26 version includes residence information
- IRS does not review projects that do not request FTI
  - Most demographic projects
  - All partner agency projects

Opportunities for Graduate Students: Make Investment Into Data Access

- Good to start early if this is for a dissertation proposal
  - Proposal development and review can take time
  - Hierarchy of review time – SSA>IRS>Census>Health
  - You owe Census promised benefits no matter if you finish research
- Work on a faculty member's proposal or existing project
  - Check in RDC administrator whether a project at UCLA is suitable

Background on Confidentiality

- Balancing the benefits of making restricted data available to the research community and the legal requirement to ensure respondent confidentiality.
- Disclosure of confidential material is prohibited by law:
  - Title 13 U.S.C. section 9 prohibits the disclosure of confidential information.
  - Disclosure is punishable by a fine of up to $250,000 or a prison term of up to five years (or both).
- Federal Tax Information (FTI):
  - Many economic datasets are “comingled” with IRS data
  - Title 26, U.S.C. Sections 7213, 7213A, and 7431 provide civil and criminal penalties for unauthorized use or disclosure of FTI.
  - Punishable by a fine of up to $250,000 or a prison term of up to five years (or both).

Special Sworn Status

- Special Sworn Status (SSS) with the Census Bureau is required to work from an RDC – regardless of which data you use.
  - SSS is granted to experts who can help the Census Bureau fulfill its mission. SSS holders are sworn for life to protect confidentiality.
- Application includes risk assessment and background check.
  - Separate from proposal review
  - 2-3 months to process (slightly longer for foreign nationals)
  - No fee if only Census data (partner agencies may charge)
  - SSS is maintained through mandatory annual trainings

Disclosure Prevention

- A number of steps are taken to limit the risk of disclosure
- Physical security limits access to the lab
  - Badged access - alarm protected lab
  - Thin Clients – no data onsite
- Special Treatment of NCHS data
  - The RDC administrator has to be on site during data access

Releasing Results

- Disclosure Avoidance Review
  - Process to review output to ensure no risk of disclosure
- Performed by RDCA and/or agency disclosure officer
  - Review process worked out in proposal stage
    - Catalog all samples, report cell sizes, detailed memos describing all releases
    - Turn around generally 1-2 weeks but plan for 3-4 weeks
    - Descriptive data can be problematic
    - Limit Intermediate output

Part 3: Some Stats About the RDCs

Public Program and Project Statistics

- Program Statistics
  - https://www.census.gov/about/adrm/fsrdc/about/stats.html

- Project Abstracts and Meta Data
  - https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html

Summary of Introduction to RDC

RDC is a growing resource for researchers to access confidential data

1. Demographic Data
2. Economic Data
3. Health Data
4. BLS Data
5. BEA Data
6. Merged Data

There is some up-front cost, but it is worthwhile to plan ahead

- Many research projects take longer anyways!

We are here to help

- RDC will have increasing resources to help in proposal writing!

Bonus Slides

Synthetic Data Alternatives

- "Synthetic" versions of some popular micro-data are available
  - Data are simulated from statistical models and designed to mimic the distributions of the underlying real data
  - Results can be verified against real data
  - Easy access; preparation for full RDC proposal
- Access through Cornell University
  - SynLBD
  - SIPP Synthetic Beta (SIPP – SSA linked)

Universe or Sample?

- Universe
  - Establishments
    - LBD, SSEL, BR, Economic Census
  - Persons
    - Census Numident
  - Workers
    - LEHD (within participating states)
  - Transactions
    - LFTTD

- Sample
  - Establishments
    - Annual economic surveys held in intercensal years
    - BRDIS/SIRD, SBO, MEPS-IC, ACES, PACE, MECS
  - Domestic Shipments
    - CFS
  - Persons
    - ACS, SIPP, CPS, NCVS, NLMS, etc.

Linking External Data to Internal Data

- External data aggregated above individual level
  - Contextualize person or establishment records with external data at tract, zip code or county level
  - Describe in proposal
- Linking on Individual level (persons)
  - Protected Identification Keys (PIKs)
  - Not all internal micro-data are PIKd
  - For external data to be PIKd:
    - SSN, name, place of birth, address, etc.
    - MOU between Census and data owner
    - Additional fee paid to Census to PIK external records
  - Clearinghouse, ERD, CLIP

Pathways for Graduate Student Access

- Work on existing project
  - Contact Administrator or Executive Director to see if existing project fits your interest
  - Work will need to fall within the scope of existing project
- Make own application
  - Start early
  - Consult with advisor