Accessing & Safeguarding Administrative Data at CCPR: Census Research Data Center (RDC)

Till von Wachter, Director, FSRDC

John Sullivan, Administrator, FSRDC

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

What is the RDC system?

- 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. Administrative Data (IRS)
- 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 Take Aways 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 not that hard

- Different data sets require different procedures
- 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 24 RDCs (and branches)
- Large number of active research projects (200+)

Goal of RDC is to make data accessible while safeguarding confidentiality

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

An Increasing Number of Agencies is Using RDC Network

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

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. **Economic Data (Census Bureau)**
   - Ex: Economic Census, Annual Survey of Manufacturing, Longitudinal Business Database (LBD), etc.
3. **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
4. **Labor Data (Bureau of Labor Statistics)**
   - Ex: NLSY with geocodes; occupation injury statistics
5. **Merged and Administrative Data Sets**
   - Ex: Longitudinal Employer Household Dynamics (LEHD)

1. Demographic Data – Why Use the RDC?

- Micro-data with detailed geography
  - Tract level for most (block-level for Decennial and ACS)
  - 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
  - 1950 - 2010
- American Community Survey (ACS)
  - Annual microdata, 1996-2015
- Current Population Survey (CPS)
  - Various Supplements (including March ASEC)

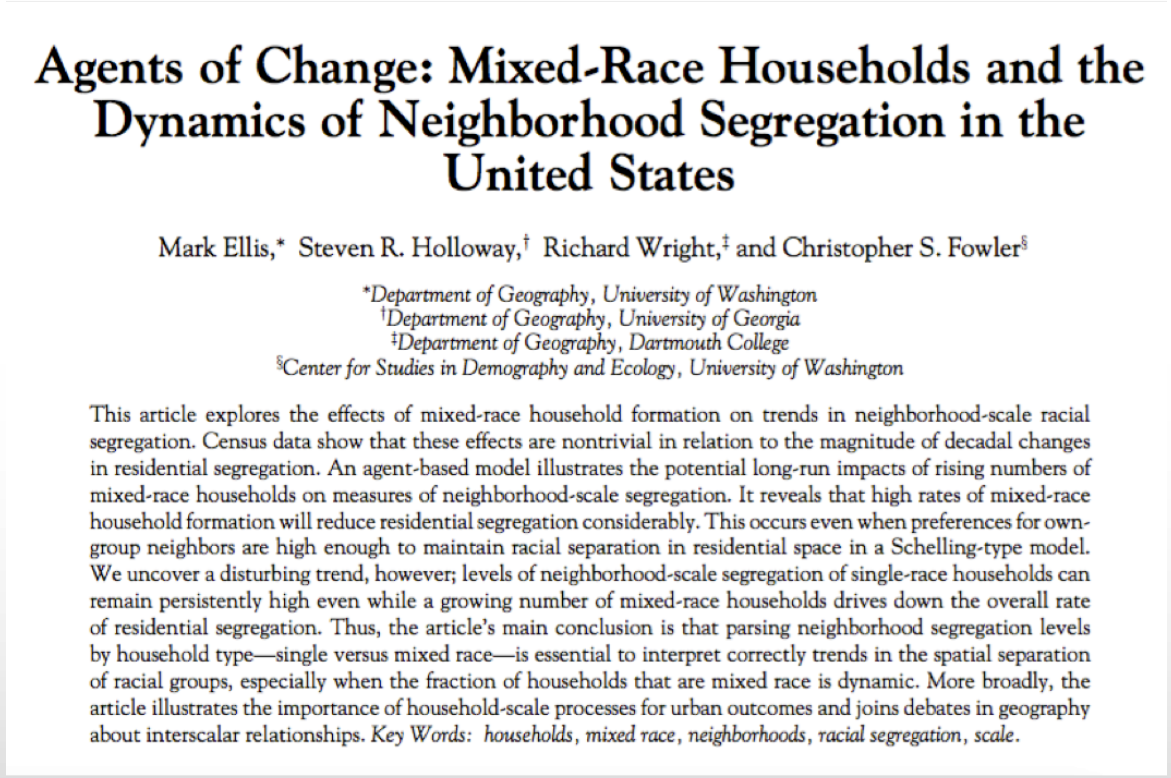
1. Demographic - Available Data (2)

- 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

1. 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

Example: Research from the UCLA RDC



Agents of Change: Mixed-Race Households and the Dynamics of Neighborhood Segregation in the United States

Mark Ellis,* Steven R. Holloway,† Richard Wright,‡ and Christopher S. Fowler§

*Department of Geography, University of Washington
†Department of Geography, University of Georgia
‡Department of Geography, Dartmouth College
§Center for Studies in Demography and Ecology, University of Washington

This article explores the effects of mixed-race household formation on trends in neighborhood-scale racial segregation. Census data show that these effects are nontrivial in relation to the magnitude of decadal changes in residential segregation. An agent-based model illustrates the potential long-run impacts of rising numbers of mixed-race households on measures of neighborhood-scale segregation. It reveals that high rates of mixed-race household formation will reduce residential segregation considerably. This occurs even when preferences for own-group neighbors are high enough to maintain racial separation in residential space in a Schelling-type model. We uncover a disturbing trend, however; levels of neighborhood-scale segregation of single-race households can remain persistently high even while a growing number of mixed-race households drives down the overall rate of residential segregation. Thus, the article's main conclusion is that parsing neighborhood segregation levels by household type—single versus mixed race—is essential to interpret correctly trends in the spatial separation of racial groups, especially when the fraction of households that are mixed race is dynamic. More broadly, the article illustrates the importance of household-scale processes for urban outcomes and joins debates in geography about interscalar relationships. *Key Words: households, mixed race, neighborhoods, racial segregation, scale.*

- Ellis, et. al.
- Decennial Census 2000 long-form data
  - Tract-level location of mixed race households
  - Mixed-race household formation reduces metropolitan level racial segregation

2. 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

2. 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

2. 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)

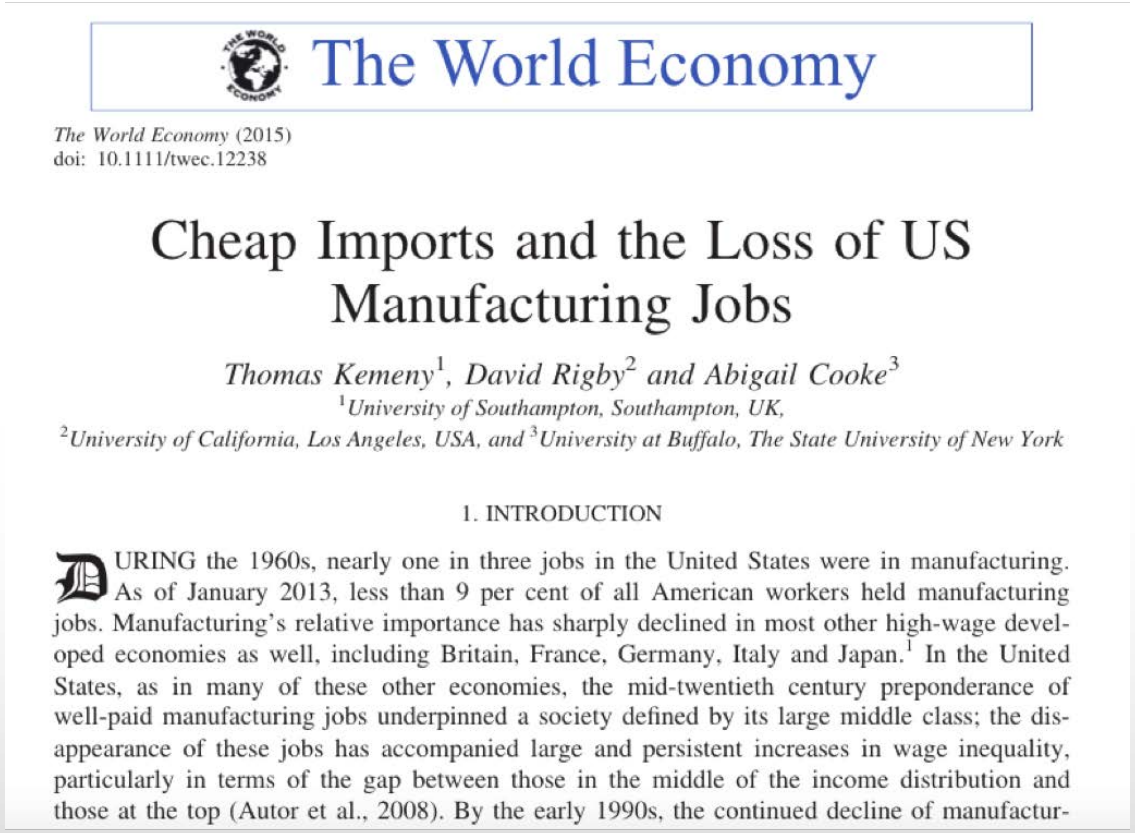
2. Business - Available Data (3)

- 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



The World Economy

The World Economy (2015)
doi: 10.1111/twec.12238

Cheap Imports and the Loss of US Manufacturing Jobs

Thomas Kemeny[1], *David Rigby*[2] *and Abigail Cooke*[3]
[1]*University of Southampton, Southampton, UK,*
[2]*University of California, Los Angeles, USA, and* [3]*University at Buffalo, The State University of New York*

1. INTRODUCTION

DURING the 1960s, nearly one in three jobs in the United States were in manufacturing. As of January 2013, less than 9 per cent of all American workers held manufacturing jobs. Manufacturing's relative importance has sharply declined in most other high-wage developed economies as well, including Britain, France, Germany, Italy and Japan.[1] In the United States, as in many of these other economies, the mid-twentieth century preponderance of well-paid manufacturing jobs underpinned a society defined by its large middle class; the disappearance of these jobs has accompanied large and persistent increases in wage inequality, particularly in terms of the gap between those in the middle of the income distribution and those at the top (Autor et al., 2008). By the early 1990s, the continued decline of manufactur-

- Kemeny, Rigby and Cooke
- LEHD linked to Census of Manufactures, FT – Import/Export, etc.
  - Rising import competition from developing economies increases likelihood of job loss among less educated U.S. workers.

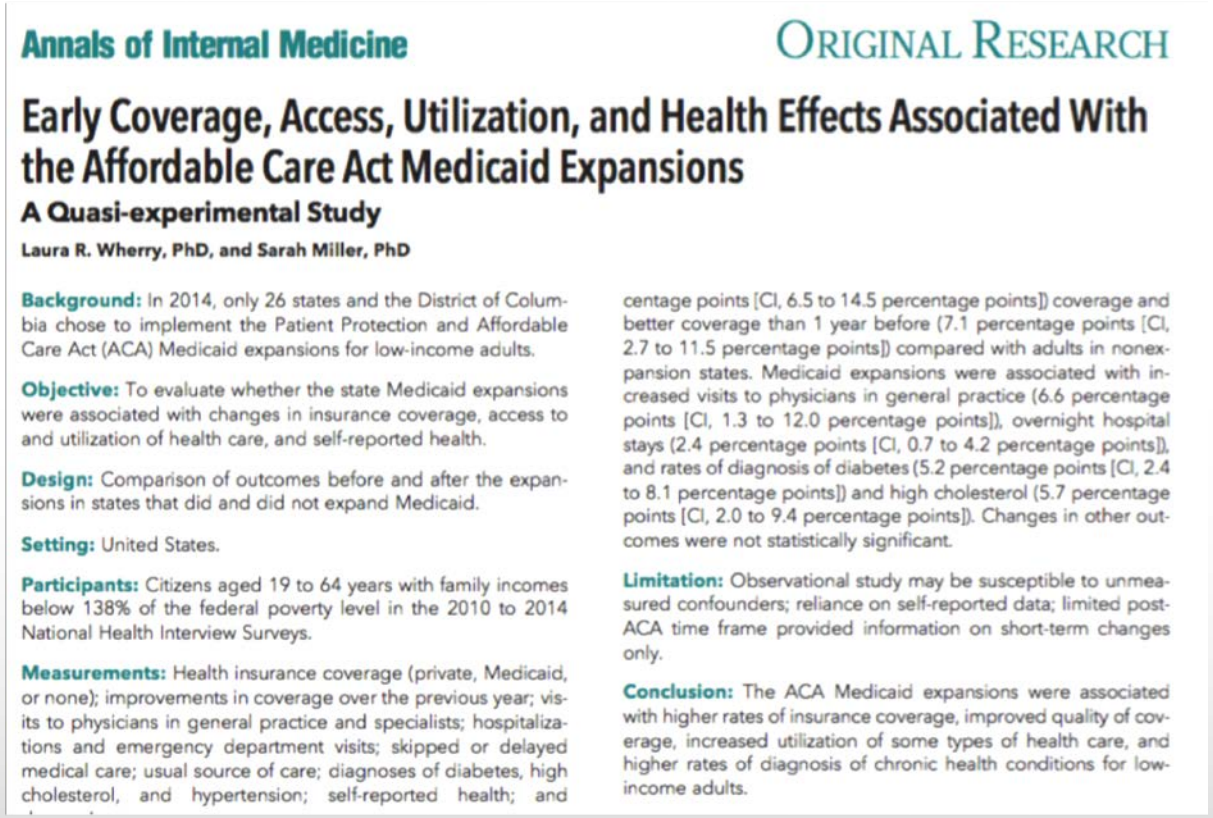
3. Health Data - Why use the RDC?

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

3. 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

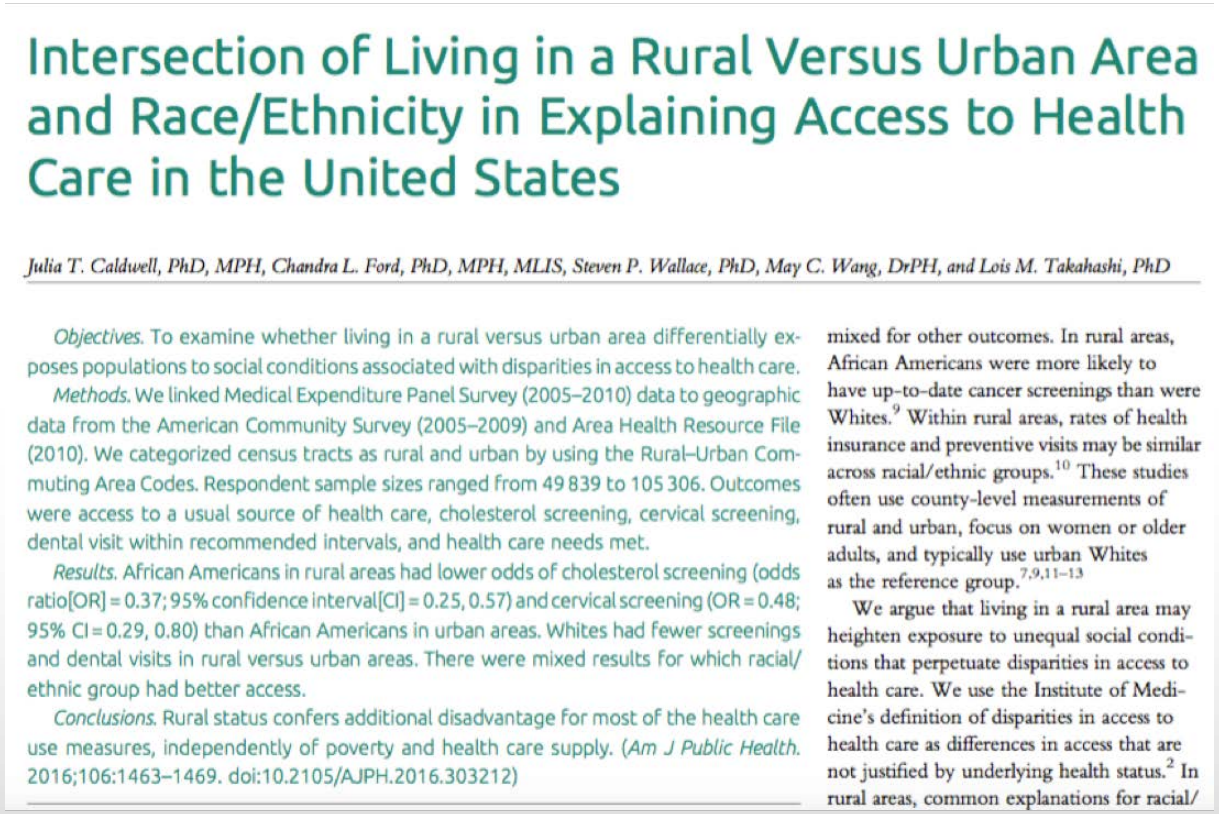
centage points [CI, 6.5 to 14.5 percentage points]) coverage and better coverage than 1 year before (7.1 percentage points [CI, 2.7 to 11.5 percentage points]) compared with adults in nonexpansion states. Medicaid expansions were associated with increased visits to physicians in general practice (6.6 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



Intersection of Living in a Rural Versus Urban Area and Race/Ethnicity in Explaining Access to Health Care in the United States

Julia T. Caldwell, PhD, MPH, Chandra L. Ford, PhD, MPH, MLIS, Steven P. Wallace, PhD, May C. Wang, DrPH, and Lois M. Takahashi, PhD

Objectives. To examine whether living in a rural versus urban area differentially exposes populations to social conditions associated with disparities in access to health care.

Methods. We linked Medical Expenditure Panel Survey (2005–2010) data to geographic data from the American Community Survey (2005–2009) and Area Health Resource File (2010). We categorized census tracts as rural and urban by using the Rural–Urban Commuting Area Codes. Respondent sample sizes ranged from 49 839 to 105 306. Outcomes were access to a usual source of health care, cholesterol screening, cervical screening, dental visit within recommended intervals, and health care needs met.

Results. African Americans in rural areas had lower odds of cholesterol screening (odds ratio[OR] = 0.37; 95% confidence interval[CI] = 0.25, 0.57) and cervical screening (OR = 0.48; 95% CI = 0.29, 0.80) than African Americans in urban areas. Whites had fewer screenings and dental visits in rural versus urban areas. There were mixed results for which racial/ethnic group had better access.

Conclusions. Rural status confers additional disadvantage for most of the health care use measures, independently of poverty and health care supply. (*Am J Public Health.* 2016;106:1463–1469. doi:10.2105/AJPH.2016.303212)

mixed for other outcomes. In rural areas, African Americans were more likely to have up-to-date cancer screenings than were Whites.[9] Within rural areas, rates of health insurance and preventive visits may be similar across racial/ethnic groups.[10] These studies often use county-level measurements of rural and urban, focus on women or older adults, and typically use urban Whites as the reference group.[7,9,11–13]

We argue that living in a rural area may heighten exposure to unequal social conditions that perpetuate disparities in access to health care. We use the Institute of Medicine's definition of disparities in access to health care as differences in access that are not justified by underlying health status.[2] In rural areas, common explanations for racial/

- Julia Caldwell, et. al.
  - MEPS-HC, ACS and Area Resource File (ARF)
  - Rural v. Urban exposure to social conditions associated with disparities in access to health care.

4. 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

5. Linkage/Administrative Data

- Possible to link external data to restricted data on the individual level
  - PIKs (Protected Identification Keys)
  - Linkage over time in mandatory collections is restricted
- CenHRS

3. Longitudinal Employer Household 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

Part 2: How To Access RDC Data

Overview: Data Access

Basic Procedure for Census 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
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, BLS:

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

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)

Application Process: Census vs. NCHS, AHRQ and BLS

- Census proposals are submitted to the RDC administrator
- Separate proposal process for NCHS/AHRQ/BLS: submit direct to agency – does not go through the RDC administrator
  - Must contact RDC administrator before submitting to NCHS/AHRQ/BLS
- 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

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 Proposals

- Plan ahead
  - At least 6 months (health) to a year (business) to get access
- Work with the RDC Administrator
- Written for a data expert rather than a content expert
  - IRS may also be a reviewer and should be considered
- Description limit is 15 pages single spaced (30 pages double)
- Benefits often emerge as proposal is developed

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 and Challenges for Graduate Students

- Proposal development and review take a significant amount of time
  - Hierarchy of review time – SSA>IRS>Census>Health
- Work environment
  - Using a server cluster; data cleaning and documentation; disclosure review; advisor access
- Good to start early if this is for a dissertation proposal
- Work on a faculty member's proposal or existing project

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
  - SSS is maintained through mandatory annual trainings

Disclosure Prevention

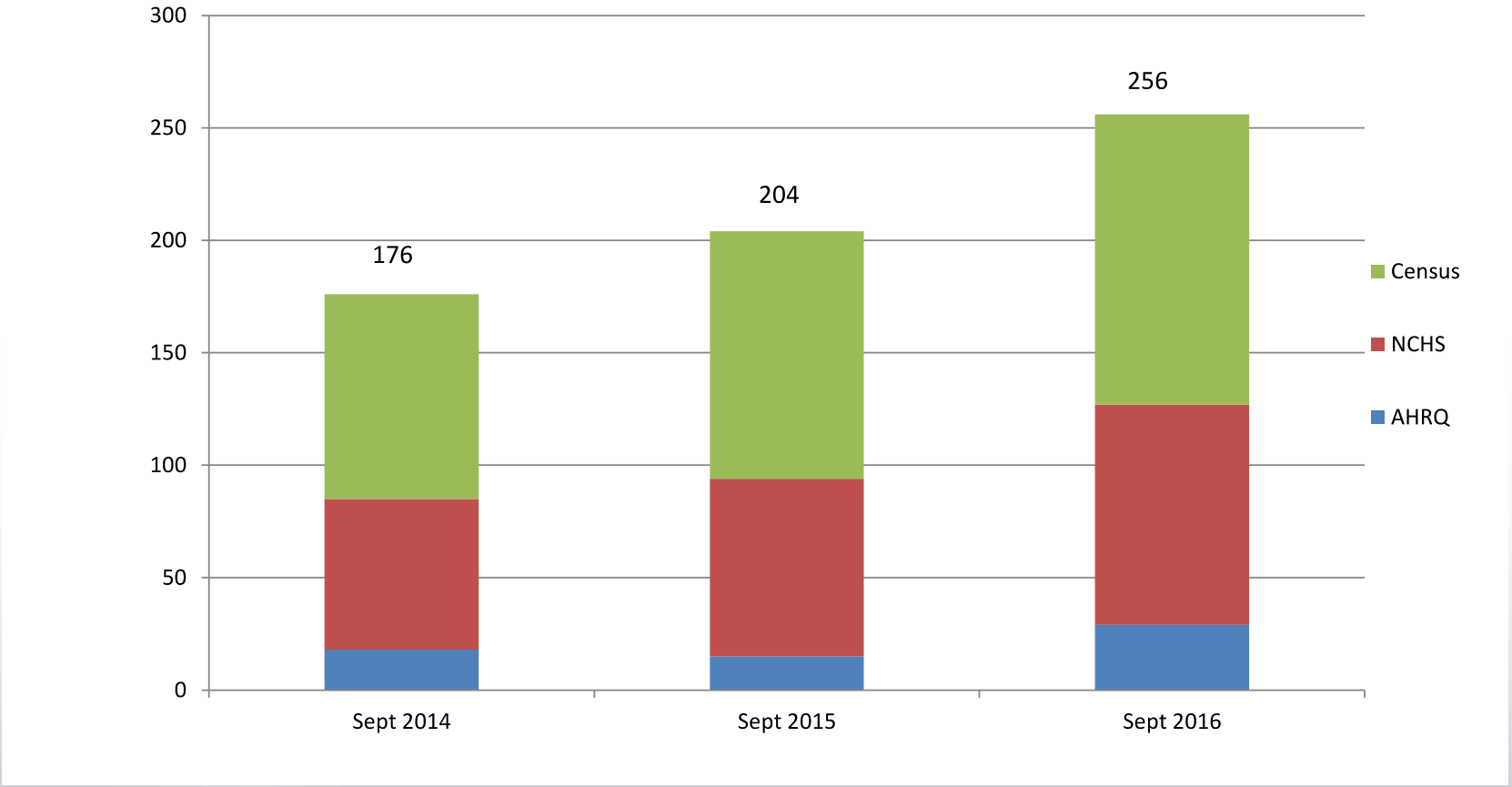
- A number of steps are taken to limit the risk of disclosure
- Physical security limits access to the lab
  - Badged access - alarm protected lab
  - Thin Clients – no data onsite
- Special Treatment of NCHS and AHRQ data
  - The RDC administrator has to be on site during data access
  - Currently,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

Releasing Results

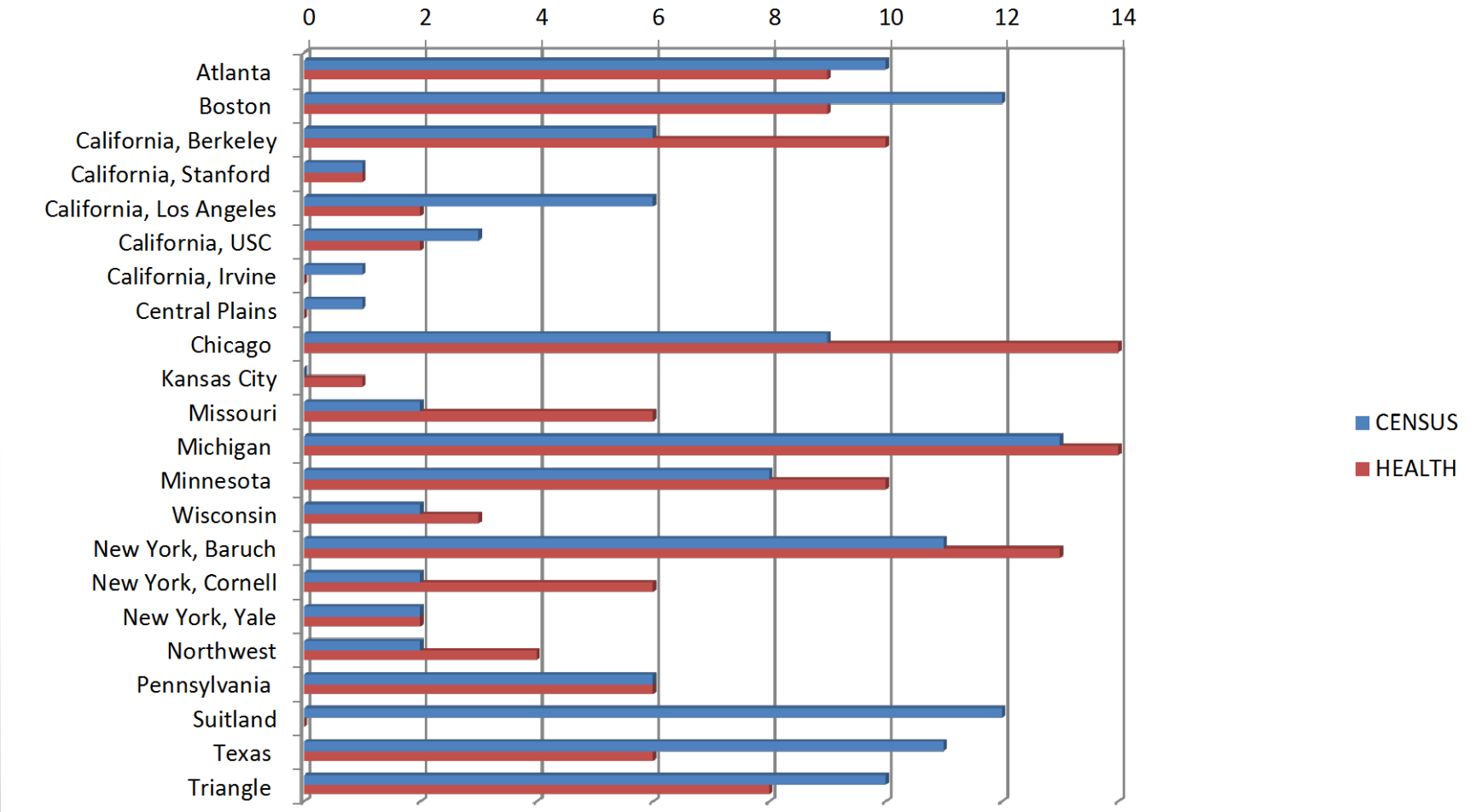
- Disclosure Avoidance Review
  - Process to review output to ensure no risk of disclosure
- Performed by RDCA and/or agency disclosure officer
  - Review process worked out in proposal stage
    - Catalog all samples, report cell sizes, detailed memos describing all releases
    - Turn around generally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

Part 3: Some Stats About the RDCs

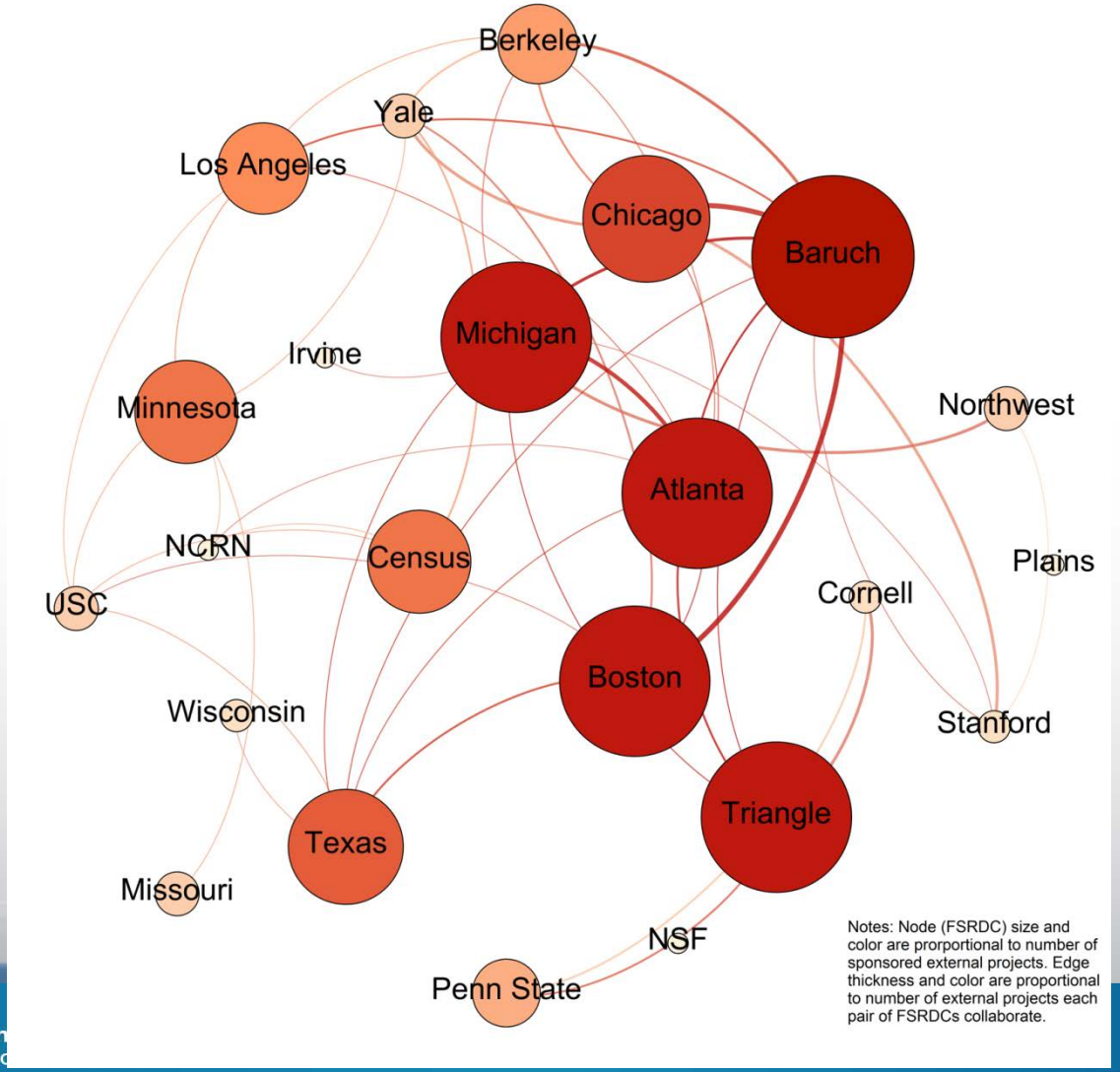
FSRDC Active Projects by Type



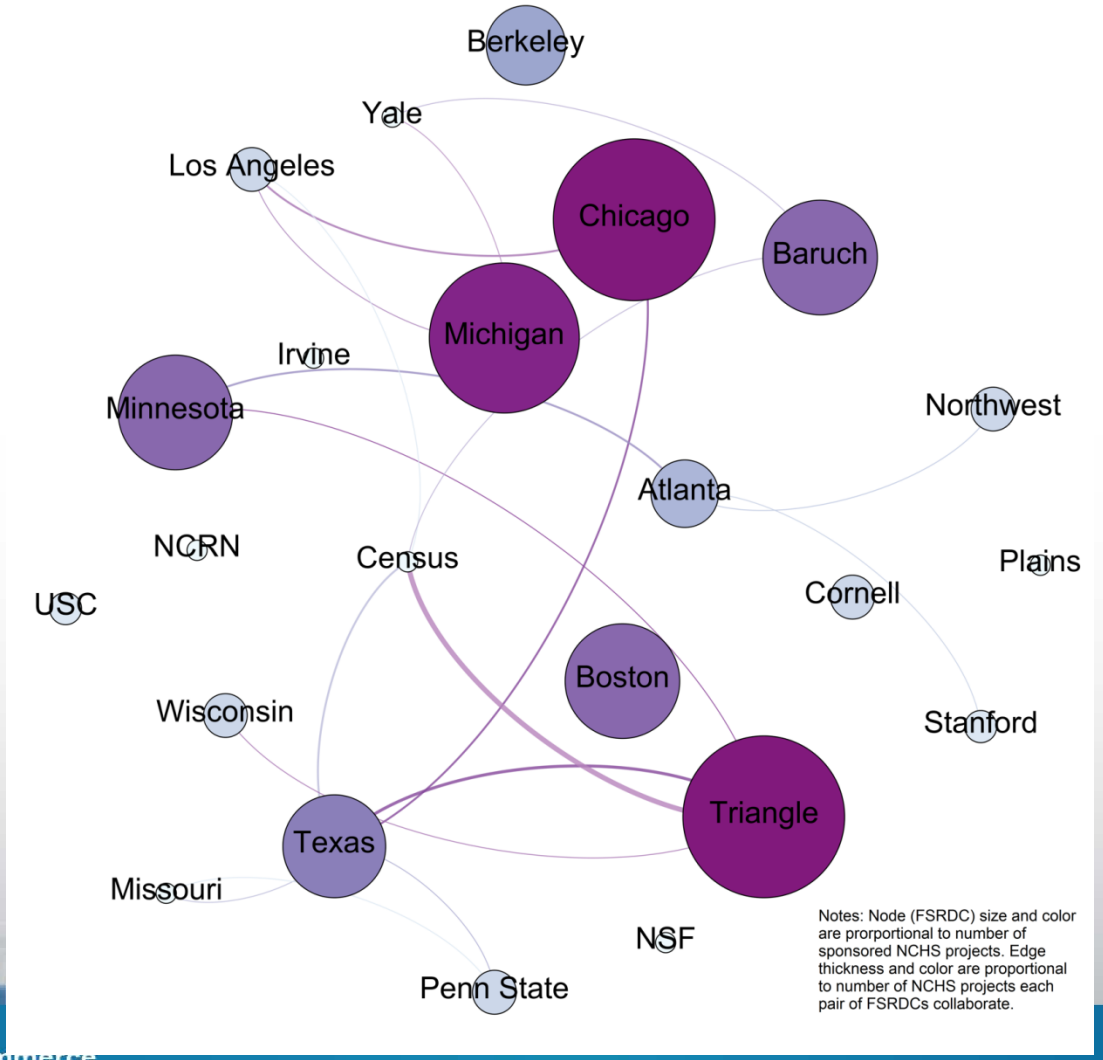
Projects by RDC and by Type



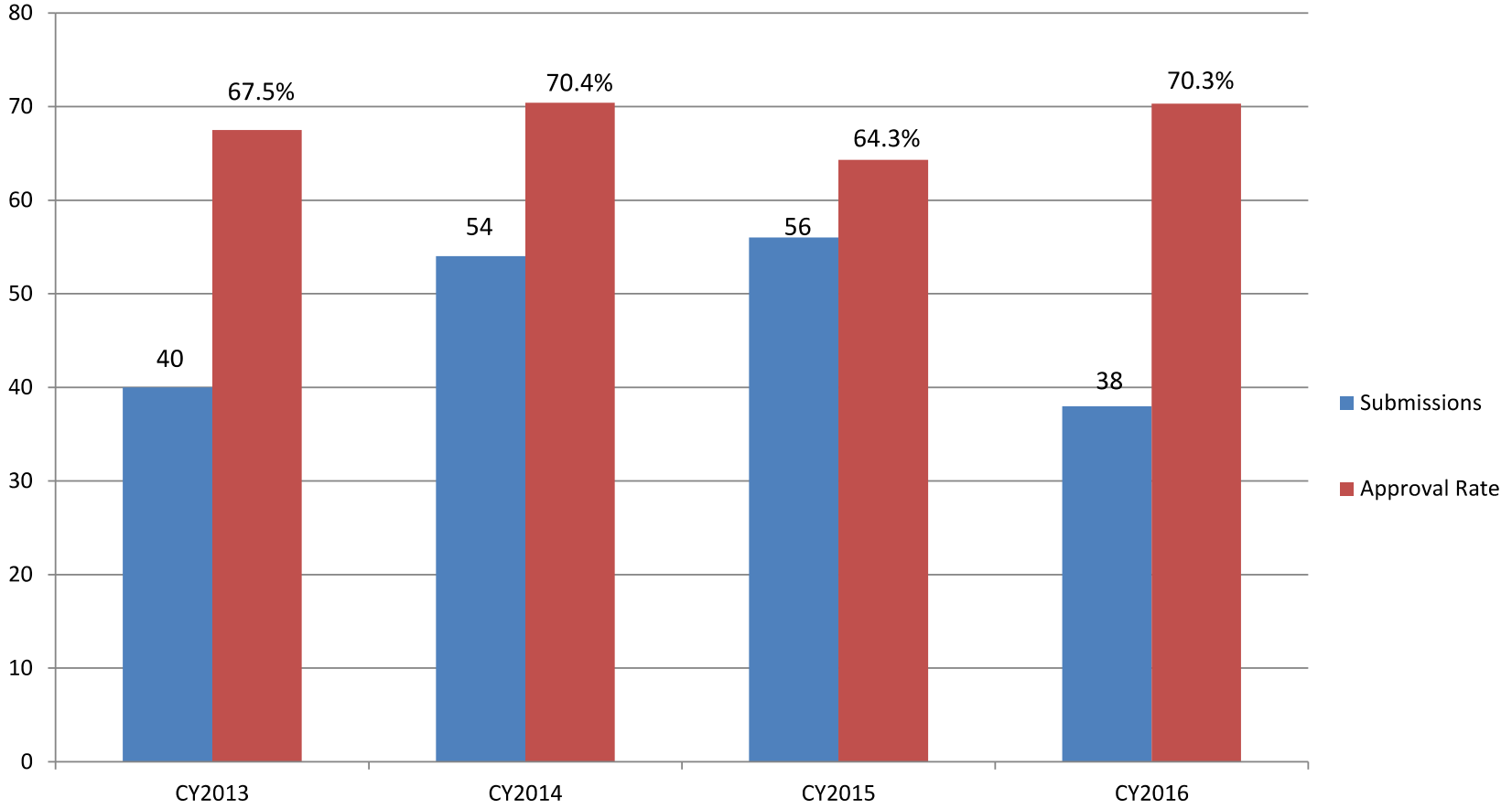
Graphical Representation – Census Projects



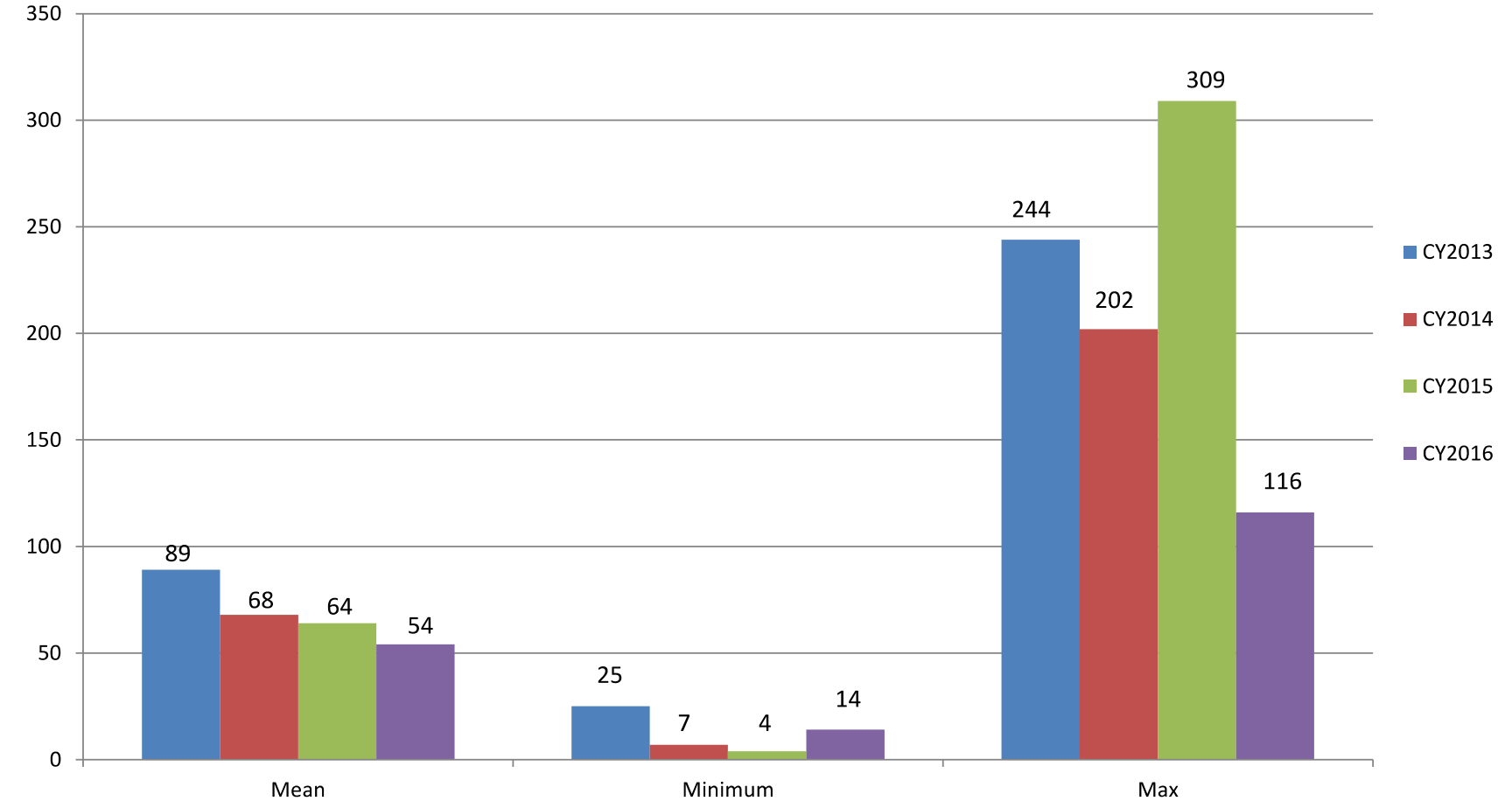
Graphical Representation – NCHS Projects



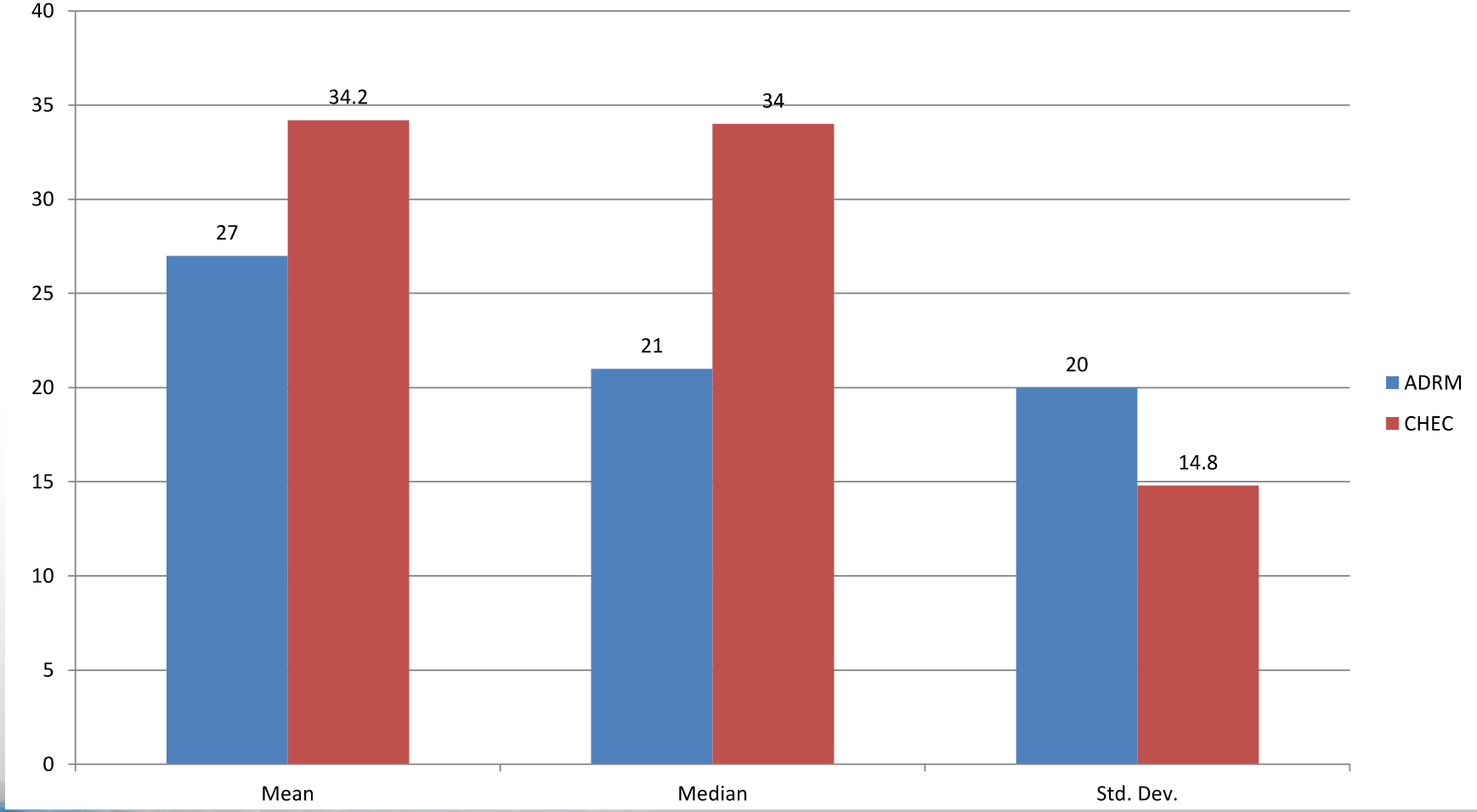
Approval Rates



Census Review Duration



Duration of SSS Processing



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. 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, CARRA, 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