Notes on Opportunities for Research Using Restricted Demographic and Health Data

Mark Fossett

Executive Director, Texas Research Data Center
College of Liberal Arts, Texas A&M University
College Station, Texas

Dallas Federal Statistical Research Data Center
Dallas Federal Reserve Bank

December 5, 2018
Overview of Presentation

Congratulations to the Dallas FSRDC and Wenhua Di

1. Overview of FSRDCs and Texas

2. Examples of Demographic & Health Data Sets

3. Examples of Demographic & Health Projects

4. Scouting RDC Projects and Data

5. Preparing Proposals for Projects Using Health Data

6. RDC Research and Careers
In 2012 the national network of RDCs consisted of 11 Core sites and 4 Branch locations sited at elite research institutions
The Texas RDC-TAMU joins the RDC network, opening fall 2012
The Texas RDC-UT Austin joins the RDC network fall 2018
What is a Federal Statistical Research Data Center (RDC)?

RDCs are unique research facilities where the U.S. Census Bureau partners with leading research institutions around the nation.

Each RDC has a *secure computing lab* that is linked with the U.S. Census Bureau’s internal computing network in Washington DC.

The Headquarters RDC computing system hosts restricted-access data sets from various federal statistical agencies.

Local RDCs are managed by an on-site Census Bureau employee – the RDC Administrator – who serves as a liaison with the research community.
The Opportunity and Conditions of RDC Research

The RDC contract enables “qualified” researchers who have approved projects to conduct statistical analyses using non-public (restricted) data from the federal statistical system.

Note that researcher access is based on several conditions.

• Researchers must be “qualified” based on professional credentials and, importantly, having “special sworn status” which makes them an (unpaid) Census Bureau researcher.

• Research projects must undergo under external agency review.

• Projects must provide benefits to agencies in keeping with regulations applying to their data.

• Data analysis in the RDC lab and release of results both follow strict protocols to protect confidentiality of the data.
Why are RDCs Needed?

Federal laws and regulations strictly protect confidential data in the federal statistical system

- It is illegal to disclose confidential federal data
- The restrictions reflect a legal, ethical, & practical concerns

Federal employees can access restricted data only as needed to fulfill their mission as outlined in statute.

RDCs extend the possibility of accessing restricted data to researchers via the mechanism of Special Sworn Status (SSS)

- In effect, the researcher becomes as an unpaid Census Bureau employee whose efforts benefit the Census Bureau
- SSS is conferred via a formal, mandatory review process
What Do RDCs Offer?

Primary Benefits and Resources

- **ACCESS**
  RDCs provide qualifying researchers access to restricted data in the federal statistical system.

- **SECURE COMPUTING LAB**
  All statistical analysis is conducted on site in the secure lab.

- **RDC ADMINISTRATOR SERVICES**
  RDC Administrators assist researchers in developing and submitting project proposals, overseeing the research environment, and assisting with disclosure avoidance reviews.
What Research Communities Do RDCs Serve Best?

RDCs Serve Well ...

- Basic Science Research
  - discipline-based substantive research
  - statistical and methodological research

- Planning and Policy Science Research
  - research evaluating program effectiveness
  - research assessing program impacts

- Research programs with long time horizons

RDCs Serve Less Well (or not at all) ...

- Research with short time horizons
- Research aiming to produce detailed lists and tabulations
- Projects aiming to produce data products for public distribution and analysis outside of the secure lab
Advantages of RDC Access to Researchers

MICRODATA
Many microdata files are not available publicly

KEY VARIABLES/MEASURES
Variables and combinations of variables not available publicly
(e.g., low level geography, sensitive health information, etc.)

LARGER DATA SETS
Full population counts and/or larger samples

INITIAL ITEM RESPONSES BEFORE PROCESSING
Access to original item responses before “editing” and processing
(e.g., detailed race answers, income is not top-coded, etc.)

MORE DETAILED CODES FOR KEY VARIABLES
Finer coding of variables (e.g., race, 5 digit industry codes)

ABILITY TO LINK/JOIN DATA SETS
Researchers can link external data (e.g., via geocodes, establishment ID, etc.) also other *internal restricted* data
RDC-Based Research is Prominent in Many Fields

Leading researchers in many fields conduct research in RDCs

- Business, Trade, Finance, and Management
- Crime and Crime Victimization
- Demography and Population Studies
- Economics, Labor Markets, Entrepreneurship, Employment and Industry
- Education and Education Policy
- Hazard Reduction, Environmental Impact Assessment, Pollution Abatement
- Health and Well-Being, Health Insurance, and Health Policy
- Housing, Housing Markets, and Residential Patterns
- Poverty and Social Welfare Policy
- Transportation Analysis and Planning
- Urban and Regional Economics and Planning
Available Data Sets Number in the 100’s

1. AHRQX MEPS Extract
2. American Community Survey
3. American Housing Survey
4. Annual Capital Expenditures Survey
5. Annual Retail Trade Survey
6. Annual Survey of Manufactures
7. Auxiliary Establishment – ES9200
8. Business Expenditures Survey
9. Business Register Bridge
10. Census of Construction Industries
11. Census of Finance, Insurance, and Real Estate
12. Census of Manufactures
13. Census of Mining
14. Census of Retail Trade
15. Census of Services
16. Census of Transportation, Communications, and Utilities
17. Census of Wholesale Trade
18. Commodity Flow Survey
19. Compustat-SSEL Bridge
20. Current Industrial Reports
Available Data Sets – Continued

22. Decennial Census Long Form Sample
23. Decennial Employer-Employee Database
24. Economic Census of Puerto Rico
25. Employer Characteristics File
26. Employment History Files
27. Enterprise Summary Report – ES9100 (large company)
28. Exporter Database
29. Foreign Trade Data – Export
30. Foreign Trade Data – Import
31. Form 5500 Bridge File
32. Geocoded Address List
33. Individual Characteristics File
34. Integrated Longitudinal Business Database
35. Longitudinal Business Database
36. Manufacturing Energy Consumption Survey
37. Medical Expenditure Panel Survey (MEPS) – Insurance Component
38. National Center for Health Statistics Data Extract
39. National Employer Survey
40. National Longitudinal Survey
Available Data Sets – Continued

41. Ownership Change Database
42. Quarterly Financial Report
43. Quarterly Survey of Plant Capacity Utilization
44. Quarterly Workforce Indicators
45. Services Annual Survey
46. Standard Statistical Establishment Listing – non Name and Address File
47. Survey of Business Owners
48. Survey of Income and Program Participation (SIPP) Panels
49. Survey of Income and Program Participation – Longitudinal
50. Survey of Industrial Research and Development
51. Survey of Manufacturing Technology
52. Survey of Plant Capacity Utilization
53. Survey of Pollution Abatement Costs and Expenditures

... and on and on and on ...

OK, you get the point
New Data Sets Become Available Regularly

New Data Sets (and versions of data sets) are added regularly

See the CES for detailed compilations
- Link to CES Census for Economic Studies (home page)
- Link to CES Notes on Restricted Data

Consult with the RDC Administrator to inquire about data sets
- Some data sets are available but are not publicized
- Important information about data sets often is not public
Health Data Sets?

Health-related research in RDCs has recently grown dramatically. Currently about 50% of RDC projects are health related, up from under 10% in 2010.

The website for the National Center for Health Statistics (NCHS) provides descriptions of the restricted variables that are available in RDCs by NCHS data set.

NCHS Restricted Data Sets & Information

The website for the US Department of Health and Human Services Agency for Healthcare Research and Quality (AHRQ) RDC provides a description of the restricted variables that are available in RDCs by AHRQ data set at the following link.

AHRQ Restricted Data Sets & Information
Part II
Scouting Restricted Data Sets and RDC Projects

1. Overview Restricted Data Sets and Restricted Variables

2. Review of Examples of Projects Using Restricted Data

3. Scouting RDC Projects
Selective Overview of Restricted Data Sets and Restricted Variables
Census Demographic and Business Data – Overview

The U.S. Census collects census and survey data in many areas including population and demography, economics and business, and more.

The Center for Economic Studies (CES) RDC website provides descriptions of data holdings and overviews of many key data sets

https://www.census.gov/ces/dataproducts/restricted_data.html
Demographic Data – Decennial Census

Decennial “Short-Form” and “Long-Form” Files

1950 - Long-form sample is 1%; Short-form sample is 100%
1960 - Long-form sample is 25%; Short-form sample is 100%
1970 - Long-form sample is 20%; Short-form sample is 100%
1980 - Long-form sample is 19%; Short-form sample is 100%
1990 - Long-form sample is 17%; Short-form sample is 100%
2000 - Long-form sample is 17%; Short-form sample is 100%

2010 – Short form only (ACS replaces Decennial Long-Form)
Demographic Data – American Community Survey

ACS Long-form files – 1% Annual Files
  All available years (2005 to present)

Note: The restricted ACS files have larger samples (approximately 1.5-2.0%)
Demographic Data – Other

Current Population Survey (CPS) & Supplements

American Housing Survey (AHS)

Survey of Income and Program Participation Panels (SIPP)

National Longitudinal Survey of Youth (NLSY)
Restricted Economic Data

Economic Data Overview

Economic Censuses:
Manufactures, Retail Trade, Wholesale Trade, Finance, Insurance and Real Estate, etc., etc.

Longitudinal Business Database (LBD) and Business Register

Annual Survey of Manufactures (ASM)

Many other firm- and establishment-level micro-data files

Commodity flow and trade micro-data files

Longitudinal Employer-Household Dynamics Data (LEHD)
Employment History Files (EHF)
Employer Characteristics File (ECF)
Examples of Restricted Census Variables

Finer levels of geography
   Lat/Lon, census block, census block group, census tract, etc.

Microdata on individuals, firms, and establishments
   Personal Identification Key (PIK)
   Employer Identification number (EIN)

Less restrictive top-coding and pre-processing
   Detailed race codes in earlier Censuses
   More detailed income data
   More detailed industry and occupation codes
More Information on Restricted Census Data

The Center for Economic Studies (CES) RDC website provides descriptions of data holdings and overviews of many key data sets

https://www.census.gov/ces/dataproducts/restricted_data.html
Restricted Health (NCHS) Data – I

National Health Status Surveys
  National Health and Nutrition Examination Survey (NHANES I, II, and III)
  National Health Interview Survey (NHIS)
  Longitudinal Study on Aging I and II (LSOA)
  National Survey of Family Growth
  National Survey of Children's Health
  National Survey of Early Childhood Health
  National Survey of Children with Special Health Care Needs
  National Survey of Children with Special Health Care Needs
  National Asthma Survey
Restricted NCHS Data – II

National Health Care Surveys

National Ambulatory Medical Care Survey
National Hospital Ambulatory Medical Care Survey
National Survey of Ambulatory Surgery
National Hospital Discharge Survey
National Nursing Home Survey (NNHS)
National Home and Hospice Care Survey
National Employer Health Insurance Survey
National Health Provider Inventory
National Immunization Survey
Restricted NCHS Data – III

Vital Statistics
  Mortality and Multiple Mortality
  Birth
  Fetal Death
  National Death Index
  Marriage and Divorce

Linked Data Sets
  Linked mortality data: NHIS, NHANES LSOA II, NNHS
  Linked Medicare Enrollment and Claims data: NHIS, NHANES, LSOA II
  Linked Social Security Administration Data: NHIS, NHANES, LSOA II, NNHS
  Linked EPA data
Examples of Restricted NCHS Variables

Geographic variables (block, block group)

Most dates (date of interview, date of death, date of birth)

Income and employment data (industry codes)

Specific diagnoses (ICD-9 codes are generally coarsened)

Details about facilities (accreditation, payments, number of employees)

Some information about children and adolescents, (e.g. height and weight, depression, behavior problems, and drug use)

Contextual data (nearest hospital, % of population with diploma)

Sample design variables (necessary for estimating variances)
More Information on Restricted NCHS Data

For further information on restricted health data see the link below

http://www.cdc.gov/rdc/index.htm
Examples of Projects Using Restricted Data
**Project Using Restricted Demographic Data**

*White & Latino Locational Attainments: Assessing the Role of Race and Resources for Segregation*

**Researchers:** Amber Crowell and Mark Fossett

**Datasets:** American Community Survey, Decennial Census

**Summary:** Objective: To assess the role of non-racial characteristics and resources in promoting/inhibiting the residential integration/segregation of Latinos. Key Findings: Latino spatial assimilation varies significantly with characteristics such as foreign-born status, English language ability, education, and income; Substantial amounts – usually half or more – of overall segregation remain even after taking account of relevant social and economic characteristics

**Reason for Using Restricted Data:** Neighborhood level geography
Project Using Restricted Demographic Data

How Low Income Neighborhoods Change: Entry, Exit and Enhancement

Researchers: Ingrid Gould Ellen and Katherine O'Regan

Datasets: American Housing Survey, Decennial Census

Summary: “This paper examines whether the economic gains experienced by low-income neighborhoods in the 1990s followed patterns of classic gentrification (as frequently assumed) – that is, through the in migration of higher income white, households, and out migration (or displacement) of the original lower income, usually minority residents, spurring racial transition in the process.”

Reason for Using Restricted Data: Tract IDs
Project Using Restricted Demographic Data

Migration and Dispersal of Hispanic and Asian Groups

Researchers: William Frey and Julie Park

Datasets: American Community Survey

Summary: “This report seeks to evaluate selective migration processes of Hispanic and Asian nationality groups in the US from established settlement areas, using recent migration data from the American Community Survey. The underlying goal is to detect migration tendencies leading toward an increased dispersion of these groups associated with their migration processes.”

Reason for Using Restricted Data: Finer geography, larger sample, detailed subgroups
Project Using Restricted Demographic Data

The Location of New Orleans Residents after Hurricane Katrina

Researchers: Narayan Sastry and Jesse Gregory

Datasets: American Community Survey

Summary: “Using individual data from the restricted version of the American Community Survey, we examined the displacement locations of pre-Katrina adult residents of New Orleans in the year after the hurricane. Over half (53%) of adults had returned to—or remained in—the New Orleans metropolitan area, with just under one third of the total returning to the dwelling in which they resided prior to Katrina.”

Reason for Using Restricted Data: Detailed geography for location of previous residence one year ago
Researchers: Carolyn Liebler and Marie DeRousse-Wu

Datasets: American Community Survey, Decennial Census

Summary: “We study the social construction of race boundaries by investigating patterns in the race, ancestry, and Mexican origin responses provided for children of 14 types of interracial marriages using dense restricted use data from 1970 to 2010. Our broader purpose is to reveal social processes that convert a newborn child of mixed heritage into an adult person of a particular race.”

Reason for Using Restricted Data: Larger sample, detailed codes for ancestry and race
Project Using Restricted Economic/Business Data

Do Environmental Regulations Disproportionately Affect Small Businesses? Evidence from the Pollution Abatement Costs and Expenditures Survey

Researchers: R. A. Becker, C. Pasurka, & R. J. Shadbegian

Datasets: Pollution Abatement Costs and Expenditures Survey, Annual Survey of Manufacturers, Census of Manufactures

Summary: “This study investigates whether the impact of environmental regulations differs by the size of the business ... We model establishments’ pollution abatement operating costs (PAOC) per unit of economic activity as a function of establishment size, industry, state, and year. Our results show that PAOC intensity increases with establishment and firm size.”

Reason for Using Restricted Data: Establishment data not public
Project Using Restricted Economic/Business Data

Capital and Labor Reallocation within Firms

Researchers: Xavier Giroud & Holger M. Mueller

Datasets: Annual Survey of Manufacturers, Census of Manufactures, SSEL/Business Register, Longitudinal Business Database, Census of Auxiliary Establishments, Airline Routes (public, linked using airline name), Compustat (public, linked using Compustat-SSEL bridge)

Summary: “We document how a plant-specific shock to investment opportunities at one plant of a company ('treated plant') spills over to other plants of the same company but only if the company is financially constrained.”

Reason for Using Restricted Data: Establishment data not public
Project Using Restricted Economic/Business Data

Do Housing Prices Reflect Environmental Health Risks? Evidence from Toxic Plant Openings and Closings

Researchers: J. Currie, L. Davis, M. Greenstone, & R. Walker

Datasets: Longitudinal Business Database, SSEL/Business Register, EPA Toxic Release Inventory (public, linked using TRI-SSEL Crosswalk), Housing Value (public, linked using geocode), Vital Statistics (public, linked using geocode)

Summary: “This paper measures the housing market and health impacts of 1,600 openings and closings of industrial plants that emit toxic pollutants ... We find that housing values within one mile decrease by 1.5 percent when plants open, and increase by 1.5 percent when plants close.”

Reason for Using Restricted Data: Establishment data not public
Project Using Restricted Economic/Business Data

University Innovation, Local Economic Growth, and Entrepreneurship

Researchers: Naomi Hausman

Datasets: Longitudinal Business Database, NSF Federal Research Funding to Universities (public, linked using geocode)

Summary: “This paper identifies the extent to which U.S. universities stimulate nearby economic activity using the interaction of a national shock to the spread of innovation from universities – the Bayh-Dole Act of 1980 – with pre-determined variation both within a university in academic strengths and across universities in federal research funding”. Finding: Employment and worker pay in relevant industries located near universities rise rapidly after Bayh-Dole.

Reason for Using Restricted Data: Establishment data not public
Project Using Restricted Health (NCHS) Data

Disparities in Diabetes:
The Nexus of Race, Poverty, and Place


Datasets: National Health and Nutrition Examination Survey, Decennial Census (public, linked using geocode)

Summary: “We sought to determine the role of neighborhood poverty and racial composition on race disparities in diabetes prevalence...We found a race - poverty- place gradient for diabetes prevalence for Blacks and poor Whites.”

Reason for Using Restricted Data: Individual health data not public
Project Using Restricted Health (NCHS) Data

Why Do They Stay? Job Tenure Among Certified Nursing Assistants in Nursing Homes


Summary: “This study identifies factors related to job tenure among certified nursing assistants (CNAs) working in nursing homes ... Among policy-relevant domains, extrinsic rewards had the largest number of significant variables.”

Reason for Using Restricted Data: Individual health data not public
Project Using Restricted Health (NCHS) Data

Racial/Ethnic Disparities in Health and Health Care among U.S. Adolescents

Researchers: May Lau, Hua Lin, & Glenn Flores

Datasets: National Survey of Children’s Health

Summary: Objective: To examine racial/ethnic disparities in health status, access to care, and use of health services in U.S. adolescents. ... Findings: Certain disparities were especially marked for specific racial/ethnic groups. Disparities included suboptimal health status and lack of a personal doctor or nurse for Latinos; suboptimal oral health and not receiving needed prescriptions for African Americans; no physician visit or mental health care in the past year for Asian/Pacific Islanders; suboptimal routine and specialty care for American Indians.

Reason for Using Restricted Data: Individual health data not public
Project Using Restricted Health (NCHS) Data

Health Disparities and Infertility: Impacts of State-Level Insurance Mandates

Researchers: Marianne Bitler, and Lucie Schmidt

Datasets: National Survey of Family Growth

Summary: Objective: “Determine whether important racial, ethnic or socioeconomic status (SES) health disparities exist in infertility, impaired fecundity, or infertility treatment. Findings: Infertility is more common for non-Hispanic black women, non-Hispanic other race women, and Hispanic women than for non-Hispanic white women, and both infertility and impaired fecundity are more common for high school dropouts and high school graduates with no college than for 4-year college graduates, and for older women compared to women under 30.”

Reason for Using Restricted Data: Individual health data not public
Project Using Restricted Health (NCHS) Data

Exposures to Fine Particulate Air Pollution and Respiratory Outcomes in Adults Using Two National Datasets: A Cross-Sectional Study

Researchers: Keeve Nachman and Jennifer Parker

Datasets: National Health Interview Survey, EPA Air Data System (public, linked using geocode)

Summary: “The objectives of this study are to evaluate the relationship between annual average ambient fine particulate matter (PM 2.5) concentrations and respiratory outcomes for adults using modeled air pollution and health outcome data and to examine PM 2.5 sensitivity across race/ethnicity... Stratified analyses revealed significant associations for non-Hispanic blacks but not for Hispanics and non-Hispanic whites”

Reason for Using Restricted Data: Individual health data not public
Scouting RDC Projects
Scouting Previous and Current Census Projects

CES Discussion Paper Series

Every Census-based project in the RDC submits at least one working paper to the CES Discussion Paper Series

The papers can be downloaded (in PDF format) at http://ideas.repec.org/s/cen/wpaper.html

The “Ideas” database is searchable

CES Annual Research Reports
Scouting Previous Health Projects

The NCHS RDC website provides a list of all publications that have come out of NCHS Projects conducted in the RDC by dataset

http://www.cdc.gov/rdc/B6Pubeyond/Pub611.htm
Practical Issues for Researches to Consider

OBTAINING PROJECT APPROVAL

- Census Bureau/Agency approval of projects is required.
- In addition to scientific merit, Census Bureau approval is based on projects providing “benefits to Census”.
- Special Sworn Status requirement for PI and all “in-lab” researchers (3-year residency requirement for non-citizens)

THE RDC COMPUTING ENVIRONMENT – Software & Hardware

- Computing is conducted in the secure lab
- The data and software are hosted on Census Bureau servers
- Census Bureau workstation hardware/software holdings

GETTING RESULTS RELEASED INVOLVES A REVIEW PROCESS

- Project results are reviewed on confidentiality criteria; NOTHING leaves the secure lab without approval
- It is crucial to plan ahead for getting project approval and for getting project output (e.g., regression results) out of the lab
Stage of Career and Planning RDC Research

Researchers at different stages of their career should think about how RDC projects fit in their plans and research agenda.

The key issues are the trade-offs between

Pro – the potential to conduct cutting edge research
Con – the time investments associated with first projects and the non-trivial time line to complete research projects

Researchers at Different Stages

Doctoral Students – must be especially careful & thoughtful
Assistant Professors – must use informed judgment
Associate and Full Professors – time to learn “new tricks” and bring junior collaborators along with you
I Want In! What Do I Do? and When?

Contact the TXRDC Director and the RDC Administrator to learn about next steps

Attend available workshops and speaker presentations

Review materials available on the web

When? Start as Soon As Possible

Census Bureau project review and special sworn status (security) review must be completed before gaining RCD access.

The process is not fast, so start as soon as possible.
Internal and External Project Review

PRELIMINARY LOCAL REVIEW

Researchers consult with the RDC Admin to learn about the review process and develop and submit proposals for agency review.

PRIMARY AGENCY REVIEW (e.g., Census, NCHS, etc.)

Typically this takes about 90 days and may involve revisions.

ADDITIONAL AGENCY REVIEW (as Appropriate)

- If non-census data are involved, additional federal agencies will be involved as appropriate.
- Relevant for IRS, Social Security, and other administrative data
- In some cases this may involve additional fees and protocols
Coordinating RDC Projects with Funding Proposals

Proposals to external funding agencies can be built around active or proposed RDC projects

- Crucial to understand how this can work so proposals can be planned and described realistically
- The RDC Director and Administrator may write support letters

External funding agencies reviews focus primarily on the project’s “scientific merit” (and fit with program goals)

RDC agency reviews focus primarily on the project’s “viability” and fit with agency needs (benefits to the agency)
Tips for a Good RDC Experience

MEET AND GET TO KNOW THE RDC DIRECTOR AND ADMIN

The Director and Admin provide guidance and explain relevant policies for access, proposals, and projects

HAVE REALISTIC EXPECTATIONS AND PLAN CAREFULLY

• Start the process early – do not underestimate the time line
• Consult early and closely to determine project feasibility

THINK STRATEGICALLY ABOUT RDC PROJECTS

When appropriate, write proposals geared to multiple papers over an extended time period (3-5 years)

LEARN ABOUT THE RDC RESEARCH ENVIRONMENT & PROTOCOLS

Attend RDC workshops; Read on-line documentation and advice; Consult with RDC researchers; Examine CES research annuals and Working Papers
Useful Sources of RDC Information

Information about Data, Data Sets, and New Developments

Center for Economic Studies (CES, US Census)
- Census Restricted Data Sets & Information

National Center for Health Statistics (NCHS)
- NCHS Restricted Data Sets & Information

Agency for Healthcare Research and Quality (AHRQ)
- AHRQ Restricted Data Sets & Information
Contact Information

Mark Fossett, Director, Texas A&M University
m-fossett@tamu.edu

Texas Federal Statistical Research Data Center
TAMU 2406
101 Donald L. Houston Building
200 Discovery Drive
College Station, Texas 77843-2406
TXRDC@tamu.edu
979-845-5618

Website: http://txrdc.tamu.edu