Using Administrative Data for Social Science Research: Promise and Peril

- Fredric C. Gey
  - UC Data Archive & Technical Assistance (UC DATA)
  - Institute for the Study of Societal Issues (ISSI)
  - University of California, Berkeley
  - http://ucdata.berkeley.edu/gey.html

- IASSIST (International Association for Social Science Information and Technology) 2010 Conference

- Cornell University, June 3, 2010
ADMINISTRATIVE DATA ARE UBIQUITOUS
i.e. What happens:

- When you get a parking ticket?
- When you apply for a driver’s license or passport?
- When you travel abroad and show your passport?
- When you take your child to school?
- When you register and vote?

- An administrative record is created!!!!

IASSIST 2010 – Using Administrative Data for Social Science Research
OVERVIEW OF THE TALK

• What are Administrative Data?
• Well-known Administrative Datasets
• Administrative Data writ small: An example of a special dataset on undocumented immigrants
• How do Administrative Data differ from Survey/Census Data?
• So you want to create your own administrative dataset
  – Collecting data from multiple sources
  – Data limitations – duplicate data; observations disappear
• Using administrative data as a basis for social surveys
  – Examples from the California Work Pays Demonstration Project
  – Missing data
  – Duplicate data
  – When people move
MY PERSONAL HISTORY WITH ADMINISTRATIVE DATA

- **1975-1977 ES-202 Establishment Data Review**
- **1979 Unemployment Insurance Claims across State Boundaries**
- **1992-1998 Work Pays Demonstration Project**
  - Collecting welfare data for California recipients
  - Evaluating welfare-to-work strategies
- **1994-2000 CalLearn Demonstration Project**
  - Evaluating incentives for teen parents to complete high school
- **2000-2002 Seasonal employment dynamics and welfare use in agricultural and rural California counties (H Brady, M Sprague)**
WHAT ARE ADMINISTRATIVE DATA 
AND HOW ARE THEY USED?

• Administrative data are data collected for the administration of government (or other) programs

• Examples include:
  – Economic data
  – Educational achievement in public schools
  – Hospital admissions/discharges/outcomes
  – Income/sales/property tax records (both personal and business)
  – Immigration applications/approvals/naturalization records
  – Social Security records
  – Unemployment Insurance claims/records
  – Voting records
  – Workers compensation (for on-the-job injuries)

• Administrative data have been used for social science research for a long time (particularly in macro-economics)
WELL-KNOWN ADMINISTRATIVE DATA SETS

• Statistics of Income (Internal Revenue Service)
  – Google scholar search on “Statistics of Income” since 2000 yields 2180 articles (6290 articles for all dates)

• BEA Employment Time Series

  On October 5, 2006, the Bureau of Economic Analysis (BEA) released estimates of full-time and part-time (FTPT) employment by industry for 1948-97 based on the 1997 North American Industry Classification System (NAICS). These employment estimates are consistent with and designed to be used with BEA's estimates of gross domestic product (GDP) by NAICS industry that were released in November 2004 and December 2005. As with the GDP by industry estimates, the employment estimates are provided for 65 detailed industries for 1977-97 and for 22 industry groups for 1948-76. With this release, BEA has completed the conversion of its industry series from the Standard Industrial Classification (SIC) system to NAICS.

• Equivalent survey (BLS Current Employment Survey)
• NCHS Mortality Detail, in particular
• Multiple Cause-of-Death Public Use Data Files
  – Derived from standard USA Death Certificate
• Basis for almost all epidemiological research on life expectancy of the United States population, e.g. accidents, homicides, suicides, alcoholism, obesity, etc.
• Problems when combining with Census Data for ethnic groups (particularly Latinos).
WELL-KNOWN DATA SETS (Continued)

• **U.S. Foreign Trade Data**
  – By Commodity
  – Imports (from other countries)
  – Exports (to other countries)
  – Example – $19 Million in horses exported to Saudi Arabia 2009
  – Example - $195,814,000 Armored Vehicles exported to Israel

• **55,200 articles/books for the Google scholar search**
  – +US +”foreign trade” +commodities
  – [+us "foreign trade" +commodities]
• **CA-237 CalWorks case load information** (monthly time series)
  – Google scholar search on “california Calworks case load” yields 872 articles

• **MediCal MEDS enrollment**
ADMINISTRATIVE DATA in the SMALL

• College Scholarship for Mexican Undocumented Immigrant High School Students in Southern California

• Small database ~200 records
  – Name
  – Date of birth → age
  – High school (e.g. San Juan Capistrano)
  – GPA
  – City and state of origin from Mexico
  – Date of application

• Duplicate names: “Maria Hernandez” from different schools

• What was missing?!!! when they came into the USA
Administrative Data VS Survey Data

• **Administrative data characteristics**
  – Restricted universe, but can have
  – Large amounts of data (millions of observations)
  – Data collected only for program administration
  – Other data spotty, even if described in program
  – Rarely includes participant opinion

• **Survey Data Characteristics**
  – Well defined sampling process
  – Small numbers of observations
    • American community survey (~200K)
    • GSS (~1500-6000) – see [http://www.du.edu/idea/director/about-gss.htm](http://www.du.edu/idea/director/about-gss.htm)
    • Public Opinion (~1200)
  – Individual opinions and characteristics often solicited (do you consider yourself Liberal/Conservative? Do you own a gun?)
UC DATA’s Experience with Work Pays

- Created (at least) 2 public use databases from welfare administrative data
- Longitudinal Database of California AFDC/CWPDP recipients from MediCal records
  - Individual records for each individual receiving aid, monthly data for eight years
- County Database for 4 California counties (Alameda, Los Angeles, San Bernardino and San Joaquin) of randomly selected welfare cases divided between control and treatment and type of case (single parent with children or two parent family)
- Used to survey for more detailed data on language, education etc, including a special language dataset for Armenian, Cambodian, Lao, Vietnamese
UC DATA’s Experience with Work Pays (continued)

- Problems encountered in database development
- Longitudinal Database of California AFDC/CWPDP recipients from MediCal records
  - Duplicate social security numbers (babies given mother’s SSN)
- County Database for 4 California counties (Alameda, Los Angeles, San Bernardino and San Joaquin)
  - Needed to rationalize data systems from 4 different counties
  - Need to understand gaps in data (people go off/on aid depending upon filling out different forms)
  - Need to understand administration rules (e.g. 2 months prospective budgeting)
  - In survey development – what information is available to create telephone interviews (phone number)?
LIMITATIONS OF ADMINISTRATIVE DATA

• Administrative data gives you numerators, not denominators if you want rates for the general population
• Confidentiality concerns abound when dealing with microdata
• Documentation and context are often poor
• For example “Statistics of Income” excludes the underground economy
The Future of Administrative Data

- Administrative data will continue to be used for public policy studies and other social research
- The major (current and) future is in record linkage across administrative programs
- For example:
  - Linking welfare/workers disability data with subsequent earnings data
  - Usually requires special access, either through a Census research center or directly from an agency
For Discussion and further reading

• Urban Institute 2007 “Catalog of Administrative Data Sources for Neighborhood Indicators”
  http://www.urban.org/UploadedPDF/411605_administrative_data_sources.pdf

• Older report (with UC DATA contribution):

• Google search “Using administrative data for social science research”