

# Technical, Legal, and Organizational Barriers to Data Linkage

Mike Larsen, Iowa State University

Thursday, October 16, 2003, 1 p.m.

NAS Workshop, Access to Research Data

# Outline

1. The Data Linkage Goal
2. Technical Barriers
3. Legal Barriers
4. Organizational Barriers
5. The Role of Statistical Modeling
6. Data Enclaves
7. HRS/SSA example
8. Concluding comments

# Themes

- Data access and data linkage are important for research
- There are barriers of various kinds to conducting linkage and accessing data
- Proactive steps are necessary to make linkage possible and enable data access while protecting confidentiality

# Data sources

- Administrative records
  - Census (Demographic, Geographic)
  - SSA (Earnings)
  - National Death Index (NCHS)
  - Program participation (Medicare, Medicaid)
- Surveys (Individuals and Households)
  - Current Population Survey
  - SIPP
  - HRS/AHEAD

# Data Linkage Goal

Utilize multiple sources to create a database with more information

- *Record linkage* – link individuals across data sources (same people in different files)
  - NDI and Employment or Medicaid, SSA
- *Combining information* – merge data based on, say, geography (different people across files)
  - CPS, SIPP, Consumer Expenditure Survey

# Example of application area

## NIA Center for Demography of Aging in Rural Communities

(Iowa State/U of Iowa applying for center)

- Populations/sample sizes are small
- Phenomena are complex:

Economic well-being, community structure and area characteristics, and health care information are in different sources

# Technical barriers: Record linkage tasks

- Parsing of fields of information (name, address)
- Standardizing fields of information
- Determining comparability of fields of info
- Measuring level of agreement/disagreement
- Forming composite weights
- Deciding which pairs of records are links
- Deciding what to do with non links

# Difficulties with record linkage

- Winkler 2003, 1995: computing challenge, processing information, string comparators
- Belin 1993: choosing cut off values critical
- Belin, Rubin 1995; Larsen, Rubin 2001: estimating error rates, modeling error rates
- Alvey, Jamerson '97: examples of challenges
- Jaro 1995: difficulties in health applications

# Further record linkage challenges

- Need programmers who can handle the job
- Field verification of linkage on a sample of cases would be great (but won't happen)
- Dates are important (need current files)
- Exact geography, good info are important
  - Larsen (1999) on residency status of administrative records not matching census

# Technical barrier: Challenges for combining info

- Geographical levels not identical
  - Multilevel modeling?
- Variables not have the same definition
- Variables recorded differently
  - Meta data? E.g., Gillman and Appel 1999
- Different survey weights
  - Good project for survey statisticians
- Statistical modeling issues ...

# Multilevel and stat'l modeling

- Multilevel modeling, hierarchical modeling
  - See Invited session, ASA SRMS, JSM 2004 Toronto
  - Levels can be states, counties, communities, households, individuals
- Small area estimation, covariates
  - Rao, 2003, *Small area estimation*
  - Fed gov experience, Fay & Herriott 1979 JASA
- Imputation for missing information
  - Clogg et al 1991 JASA, I/O codes
  - Elliott & Little 2000 JASA, census undercount

# Legal barrier: confidentiality

Laws require the identity of individuals who have information recorded in administrative databases to be kept secret from the public and other agencies

Similar promises made for surveys

This includes protection from direct, indirect or inferential, and accidental exposure

# Disclosure limitation: how?

- Severely restrict access to data
  - Obvious negative impact on research
- Restrict what can be published
  - How do you tell what can be published?
  - Who monitors? What is policy?
  - What is impact on research? How negative is it?
- Mask data to protect identity
  - Merge cells, merge geographical units
  - Add noise
    - Work by Karr, Dobra, Sanil, Fienberg and others

# Disclosure limitation, cont.

- Remote access through a security system
  - E.g., Reiter 2003, *Statistics & Computing*
  - Hard stat'l problem; combine with other protections?
- Restricted use data enclave
  - Future plans for ISU NIA demography center
  - Substantial start-up costs; potential benefits
  - Michigan documented their efforts
    - Nolte & Keller 2001, ASA SRMS Proceedings
  - More later ...

# Organizational barriers

- Secondary uses prohibited
  - Allowable secondary uses not defined?
- Strict requirements to protect confidentiality
  - Only very restrictive uses allowed
  - Protocols for sharing data not in place
- Data not ready for sharing
  - Formats non standard (what is standard?)
  - Variables defined and recorded in ‘unique’ ways (meta data issue)

# Organizational barriers, cont.

- Protecting the agency
  - Is the agency at risk from potential disclosure?
  - Is there a public relations concern?
  - Disclosure limitation must be very secure
- Agency territory
  - Could an agency prefer that other people not answer certain questions?
  - Data sets are modified over time.  
Do agencies document changes/versions sufficiently?

# Suggestions for org. barriers

- Set policy, study options
- Negotiate with other agencies if needed
- Be open to researcher suggestions
- *Anticipate need for respondent permission*
- *Standardize variables if at all possible*
  - Synchronize geography
  - Record date/time of data collection
  - Meta data repositories for surveys, admin recs

# Data Enclave, what is it?

- Physical site with offices, computers, staff, statistical software, security
- Computers hold restricted use and public use data, possibly pre processed (e.g., rural, elderly)
- Data inclusion is negotiated with agencies
- Researchers are evaluated for access approval
  - Agency funding could lead to approval
  - Proposal for specific research project

# Data enclave needs

- Need to negotiate data access
- Need to set up enclave: time, resources, space, staff, computing and statistical expertise, \$
- Need to seek continued funding
- Need protocols for approval of data access
- Need working disclosure limitation method

ISU Demography of Aging center plans to investigate data enclave option (multiyear)

# Data enclave advantages/disadv

- Databases are more usable
  - Statistical software available (several options)
  - Data potentially preprocessed (elderly, rural)
  - Consulting available?
- Interdisciplinary research opportunity
- Cross institution research opportunity

But, researchers have to travel to site and spend some time there.

Addition: secure remote access system (possible?)

# HRS-SSA linkage example

## *Health and Retirement Study*

Sponsored by NIA, NIH

Conducted by U of Michigan, SRC of ISR

n>22,000, panel every two years, 50+

Cohorts in 1992, 1993, 1998, ...

## *Social Security Administration*

65,000 employees, 1300 local offices

Contributor/beneficiary files (n=millions)

# HRS/SSA Data

## HRS Data

economics, health, demography of retirement and aging, death (J. Human Resources, v 30, 1995 supplement)

## SSA Data

earnings histories, Social Security benefit histories, Supplemental Security Income payment histories, 1992 longitudinal file

# Public versus Restricted use data

- Public use data HRS
  - Michigan, RAND (processed data)
- Public use data SSA
  - SSI recipients by county and demographic category
  - On-line tabular data
  - Micro data link suspended? (web link broken)
- Restricted use data HRS
  - Several files (see Michigan data enclave)
- Restricted use data SSA
  - Primarily internal? Or restrictive agreements

# Record linkage and HRS

- HRS/SSA – linkage of individuals
  - (Olson 1999 SSB, 62, 2, 73-84)
- HRS asks for respondent consent to link by SSN
  - (Olson 1996 SSB, 59, 1, 85-88)
- HRS recognized benefits from linkages from start
  - NDI of NCHS (Olson 1996)
  - Medicare from CMS/HCFA
  - Employer pension, health insurance plans (JHR 1995)

# Confidentiality in HRS

- Retirement History Survey (RHS), 1969-1979, cannot reinterview in 1980s due to confidentiality protections (title 13)
  - Juster & Suzman, J. of Human Resources, v30
- Concerns about census conducting survey
  - Title 13 potential limitations
- Permissions for linkages by SSN

# An SSA Linkage example

MINT: Modeling retirement income in the near term

- Census Bureau's SIPP
- SSA administrative records 1951-1996
  - Benefits, data of death

*Advantage:* combine marital and earnings histories (Butrica & Iams 1999 SSB v 62)

# Conclusions: Need for linkage

- Record linkage builds better databases
- Combining information important, especially for statistical modeling
- As example, Center for Demography of aging in rural communities at ISU (proposed) will need to combine information geographically and possibly link files to accomplish goals

# Conclusions: Challenges

- Technical
  - Record linkage challenges
  - Combining information challenges
  - Statistical modeling very important in use of info.
- Legal
  - Confidentiality critical, seek respondent permission
  - Disclosure limitation research advancing, but ...
- Organizational
  - Data enclaves? Considering supporting them
  - Study options, set policy. Access/linkage mission?

# Themes, revisited

- Data access and data linkage are important for research
- There are barriers of various kinds to conducting linkage and accessing data
- *Proactive steps* are necessary to make linkage possible and enable data access while protecting confidentiality

# Some References (please email for specific)

## Record Linkage

Gomatam, Larsen 2003, Chance  
Larsen, Rubin 2001, JASA  
Alvey, Jamerson 1997, Commerce  
Jaro 1995, Statistics in Medicine  
Belin 1993, Survey Methodology

Lahiri, Larsen 2003, JASA  
Winkler 2003, J of Information  
Belin, Rubin 1995, JASA  
Fellegi, Sunter 1969, JASA  
Winkler 1995, Business Survey

## Disclosure Limitation

Dobra, Fienberg 2000, Proceedings NAS  
Karr, Dobra, Sanil 2003, Comm. ACM

Fienberg 2001, Stats in Med.  
Reiter 2003, Stats & Computing

## Data Enclave

Nolte, Keller 2001, SRMS ASA

## Meta Data

Gillman, Appel 1999, Proceedings

# Thanks and contact

- Email: [larsen @ iastate.edu](mailto:larsen@iastate.edu) (no spaces)
- [http:// www. public. iastate. edu/ ~larsen](http://www.public.iastate.edu/~larsen)
- Thanks to Chris Mackie for invitation
- Thanks to Sarah Nusser, Taps Maiti for discussions on combining information
- Thanks to Dan Gillman for presentation at ISU on Meta Data Repositories
- Thanks to Laura Zayatz for planned comments