SHADAC

STATE HEALTH ACCESS DATA ASSISTANCE CENTER

Workshop on Data Linkages to Improve Health Outcomes

NATIONAL COMMITTEE ON VITAL AND HEALTH STATISTICS

SUBCOMMITTEE ON POPULATIONS

September 19, 2006

Michael Davern, Ph.D.

Assistant Professor, Research Director
SHADAC, Health Policy & Management
University of Minnesota

Supported by a grant from The Robert Wood Johnson Foundation

Thinking about data linkage from a data quality perspective

- Concerns with survey data for health research
- How do administrative data compare?
- Issues in merging the two sources of data
- Work left to do to fulfill great potential
  - I will not talk about data stewardship, privacy or confidentiality as it has been covered very well elsewhere

SHADAC
State Health Access Data Assistance Center

Start at the end

- There is great potential health research to be done with linked surveys and administrative data
- Survey microdata are in the public domain
  - Strengths and (especially) limitations are well known
- Administrative data are the standard on programatic issues, but….
  - Because these data are not in the public domain its imperative that limitations be thoroughly investigated by agencies entrusted with the data
  - Documentation and research on linked files (i.e. metadata) must be put into the public domain
  - NCHS, Census, NCI, SSA AHRQ have the agreements in place and are the logical people to produce this work

SHADAC
State Health Access Data Assistance Center

Survey data have well known limitations

- Survey data concerns
  - Sample frame coverage error
  - Sampling error and variance estimation
  - Non-response error (item and unit)
  - Measurement error
    - Including mixed-modes of data collection
  - Data processing, imputation/editing
  - Need for better documentation (metadata)
- In general: We know survey data have problems
  - This is a good thing!
- How are administrative data like/unlike survey data with respect to these issues?
  - Great variety of “administrative data”

SHADAC
State Health Access Data Assistance Center

Sample frame and sampling error

- Sample frame and frame coverage
  - The administrative data cover the entire enrolled population
    - Survey data link is necessary to understand the potentially enrolled population
  - When administrative data are used as a survey sample frame be careful of systematically missing contact information
- Sampling error
  - Not a problem for administrative data because it is a complete list of the enrolled population
  - Could potentially be an issue if sample drawn from administrative records and the sample is used for research

SHADAC
State Health Access Data Assistance Center

Item non-response and missing data

- Non-response error (or missing data)
  - Item non-response can be a major issue
    - Important data for research can be missing (e.g., age other program codes, or race/ethnicity).
    - Some of this data can be missing systematically
      - TANF flag by county, or race/ethnicity by state in MSIS
    - Identifying data can also be missing systematically
      - Can be a large source of sample loss for merged survey and administrative data
  - Bottom line: Administrative data have important information for health research missing

SHADAC
State Health Access Data Assistance Center

Measurement error

- Measurement error
  - Administrative data are the standard for knowing whether someone is enrolled in a program or how much someone received in benefits
  - However, other administrative data desired for research may have significant measurement error associated with it
  - Administrative data can be collected through a many modes during more than one wave of interviewing, with several instruments
    - Interviewer assisted, self-administered, completely filled out by interviewer enrollees signs, or data generated from other administrative data and added
      - Interviewers have a wide variety of training/skills (e.g. Tax accountants)
      - Medicaid enrollment data can be drawn from a wide variety of sources at the county level within states, and differently by all the states

SHADAC
State Health Access Data Assistance Center

Measurement error (continued)

- Administrative data forms are generally not as user friendly as survey self-administered forms are to fill out
- Research is needed into possible mode effects, longitudinal panel conditioning, interviewer effects, and instrumentation effects
  - Survey research has long history of this kind of work that administrative data could benefit from
- Important to remember that respondents in administrative data files may have different incentives for filling out administrative data versus survey data
  - How do these motivations lead to measurement error?
  - Also if data are not accepted unless filled out, are elements ever "curb-stoned" by interviewers or data entry folks?
- Research on data quality and measurement is essential

SHADAC
STATE HEALTH ACCESS DATA ASSISTANCE CENTER

Data editing and imputation

- Data editing and imputation and documentation
  - There is little documentation in the public domain regarding collection, editing and/or imputation procedures of administrative and enrollment data relative to survey data
  - Data editing and imputation activity happens but researchers who use the administrative data files can be caught off guard by it
  - Demographic data taken at enrollment could be imputed for research purposes using linked data
    - Something the linking of survey data with administrative data could vastly improve

SHADAC
State Health Access Data Assistance Center

How do administrative data compare to survey data for research purposes?

- Survey microdata, documentation and research into critical sources of error are all in the public domain
  - Survey data are very strong because there are so many known problems with the data
  - Similar research into problems with administrative data needs to be done
    - Especially since they are not created for the purpose of research as survey data are
    - We should not assume the administrative data have fewer of errors
    - Even though the microdata itself cannot be made public, research and documentation (metadata) used to produce the file needs to be made public if the data are to be useful for research
  - Quality of administrative data will likely vary greatly from centralized collection (IRS, SSA, Medicare) to state or county based (Medicaid and other state based)

Issues in linking survey and administrative data

- Two biggest problems in linking:
  - Universe issues
  - Measurement error
    - Both survey and administrative data
- Essential to understand differences and concordance between data sources

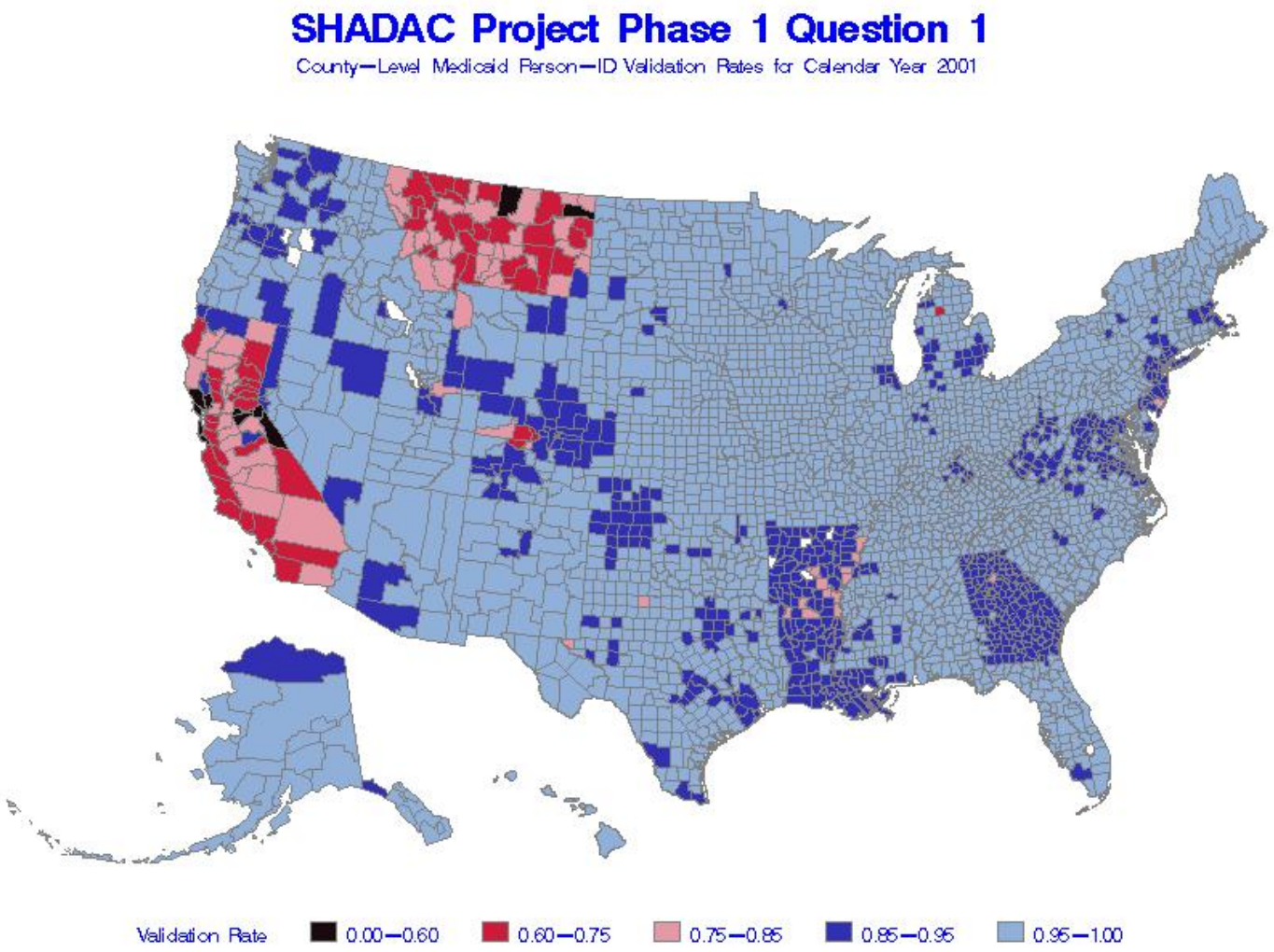
SHADAC
State Health Access Data Assistance Center

Universe Issues

- Need to understand sample loss in linked files
  - Missing linking information
    - Surveys in which respondents do not give consent, and where the linking information can not be validated
      - Differential sample loss?
    - Administrative data that is missing key linking information
      - Differential sample loss?
    - Need to build a common universe carefully

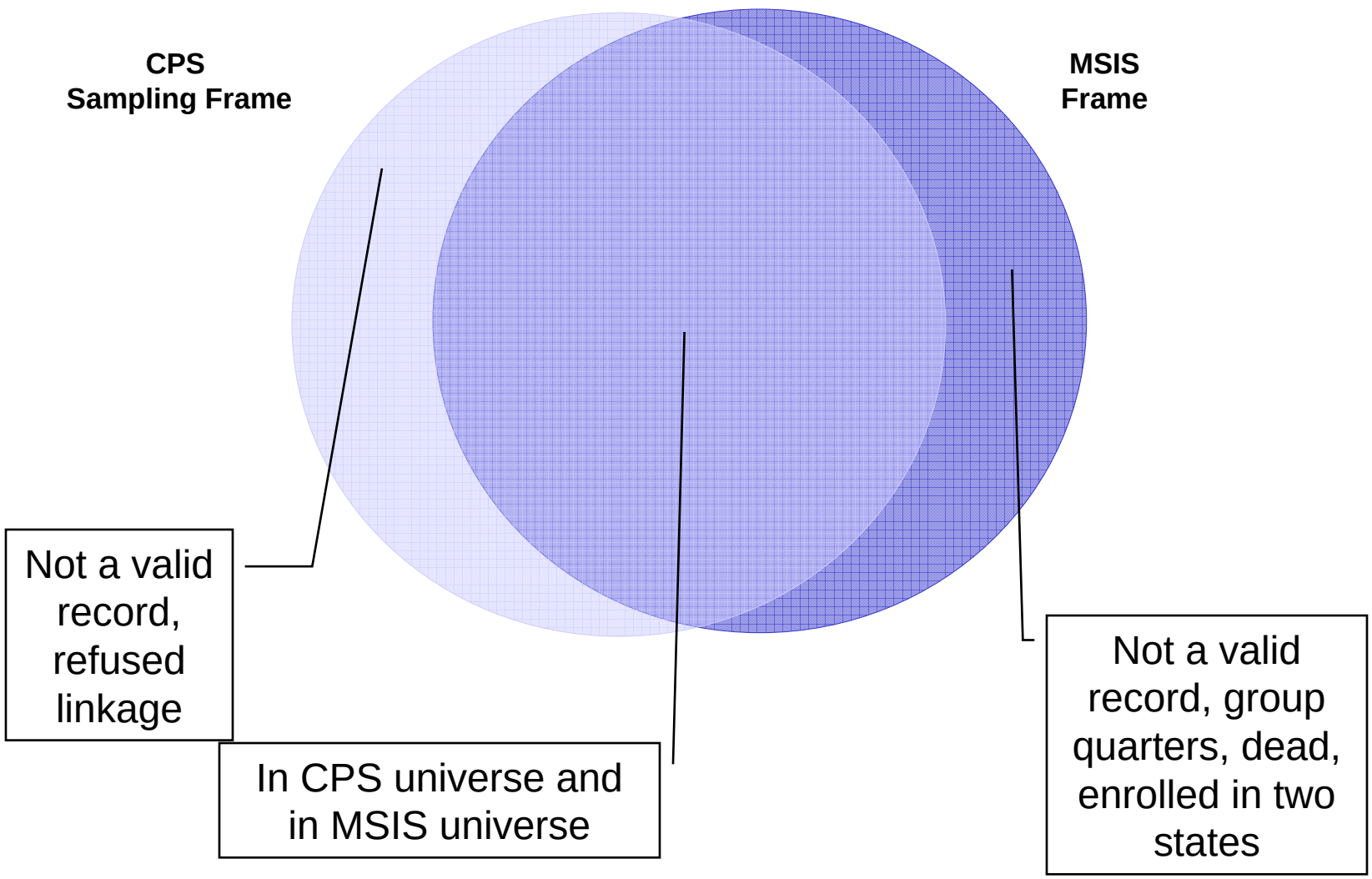
SHADAC
State Health Access Data Assistance Center

Validated Records in MSIS



SHADAC
State Health Access Data Assistance Center

Developing a linked universe



Measurement error

- Conceptual differences:
  - For example a person can be on Medicaid but not be receiving full benefits
    - Is it health insurance?
- Mis-reporting in surveys:
  - Person is on Medicaid but reports some other type of coverage or being uninsured
- Mis-classification in administrative data
  - Race data are often missing from Medicaid and are important for disparities work
  - Systematically missing variables (TANF flag, SSN in MSIS) or address and phone number in VA example

SHADAC
State Health Access Data Assistance Center

Great potential for linked data

- Potential for merged data:
  - Improving accuracy of survey data collection of enrollment data (Medicaid, SSI, etc.)
  - Improve survey sample frames (Census MAF)
  - Using merged data to create small area estimates
  - Improve administrative data race/ethnicity information
  - Great benefit to using information in imputation models and editing
  - Greatly improve health policy simulation by allowing researchers to better engage errors and appropriately model them

SHADAC
State Health Access Data Assistance Center

Wrapping up

- Problems
  - Essential data stewardship, agreements, confidentiality, privacy concerns cause problems for researchers (while doing their duty)
    - Recency: Most recent linked MSIS data is 2002
    - Data are not in the public domain and the ability to conduct research into quality of administrative data for research purposes is very limited
  - Be careful of reaching conclusions based on "asymmetrical verification"
    - For example, comparing Medicaid enrollees to the CPS to see how respondents answer the survey question only allows us to be certain if a person is enrolled and did not report enrollment or answered the survey as though they are uninsured
    - Does not allow us to verify if uninsured people report coverage
    - Making adjustments to overall uninsured based off this data alone with be one-directional
    - We are not even certain if a person who reports Medicaid in the survey, but is not matched into the administrative data, falsely claimed to have Medicaid
      - Because of sample loss
  - Sample loss can cause some strange findings if not accounted for

SHADAC
State Health Access Data Assistance Center

Last slide! -- I promise

- Bottom line:
  - Linked administrative and survey data will be very useful source of health data
  - Need to treat administrative data like survey data and examine measurement error, produce public domain documentation
    - Perhaps even meeting Data Documentation Initiative –DDI- standards
  - Research into sample loss on of linked data sets.
    - Who is missing and why?
  - Understand measurement error in survey data and administrative data.

SHADAC
State Health Access Data Assistance Center

SHADAC contact information

www.shadac.org

State Health Access Data Assistance Center
University of Minnesota
2221 University Avenue, Suite 345
Minneapolis Minnesota 55414
(612) 624-4802

SHADAC
STATE HEALTH ACCESS DATA ASSISTANCE CENTER