To Link or Not to Link? Assessing the Quality of Administrative Data for Survey Research

Joe Sakshaug
Program in Survey Methodology
Institute for Social Research
University of Michigan
August 18, 2011

Outline

- Advantages of administrative data linkage
- Potential issues with linkage
- Study 1: Do biases exist in linked administrative data?
- Study 2: Are administrative records more accurate than survey self-reports?
- Conclusions
- Future Research

Advantages of Administrative Data Linkage

Substantive research

- Obtain population-based inferences for key survey *and* administrative variables of interest
- Address complex policy-oriented research questions
  - e.g. health reform, federal assistance programs

Survey research

- Reduction in respondent burden
- Reduction in data collection costs
- Assessment of data quality

Potential Issues

1) Non-Consent

- Many surveys require respondent consent to link survey and administrative records
- Respondent consent is not universal
  - Range: 19.0% - 96.5% (McCarthy et al., 1999;Rhoades and Fung, 2004)
- Common correlates of consent (survey data)
  - Age, race/ethnicity, gender, education, marital status, health status, employment (Bates and Pascale, 2006; Jenkins et al., 2006; Banks et al., 2005; Dunn et al., 2003; Young et al., 2001; Woolf et al, 2000; Olson, 1999; Pullen et al., 1992)
  - Item missing data, interviewer characteristics, prior-wave outcomes (Sakshaug et al., 2010; Jenkins et al., 2006)
- Concern: non-consent bias in survey *and* administrative estimates

Potential Issues (cont.)

2) Accuracy of the Administrative Data

- Validity of administrative data is unknown
  - no gold standard
- Admin data can be collected from various sources with varying levels of quality
  - Population registry, employee records, credit records
  - timeliness, item missing data, noncoverage
- Some administrative data sets not designed for research purposes (e.g., billing records)
- Linking survey and admin data may yield conflicting measures of same construct (McAlpine et al., 2007; Davern et al., 2008)
  - Which measure is closer to the "truth?"

Research Questions

Study 1:

- Do non-consent biases exist for administrative data estimates?
  - Unclear; admin records typically unavailable for non-consenting cases
- What is the relative trade-off between non-consent error and traditional survey errors (e.g., NR, ME)?
  - Is it better, from a total survey error perspective, to link to admin records or ask Rs to report the same information?

Study 2:

- How accurate are administrative data compared to survey data?

Study 1: German PASS Study

- Panel Study 'Labour Market and Social Security' (PASS)
- 2006/2007 (Wave 1); RR1: 26.7%
- Mixed-mode study; CATI results shown
- Sample of benefit recipients (*Unemployment Benefit II*)
- Consent to link employment/benefit records
  - Consent requested early in questionnaire
  - 80% consent rate
- Administrative records available for all respondents and nonrespondents (consenters and non-consenters)
  - Key variables: age, nationality, employment status, monthly wage, benefit receipt, and disability status.

Verbal Consent Request

- [P23a] *"To keep the interview as brief as possible…the [IAB] could merge the study results with data about your employment, unemployment or participation in measures by the employment office."*

- *"…this cannot be done without your agreement, which we kindly ask you to provide...all rules of data protection and of the de-personalization of the results reported apply to these additional data as well."*

Bias Estimation

- Non-Consent bias (administrative estimates)
  - Consent indicator linked to administrative data

$$\bar{y}_{nc\ bias} = \bar{y}_{consenters} - \bar{y}_{resps}$$

- Nonresponse bias
  - Paradata (e.g., disposition codes) linked to admin data

$$\bar{y}_{nr\ bias} = \bar{y}_{resps} - \bar{y}_{sample}$$

- Measurement error bias
  - Two versions of same statistic obtained from PASS and administrative data

$$\bar{y}_{me\ bias} = \bar{y}_{resps,PASS} - \bar{y}_{resps,admin}$$

Bias Estimates

Variable	Non-Consent	Nonresponse	Measurement
Age	-0.3*	4.6***	-0.4
Foreign (%)	-0.9***	-5.6***	-2.5***
UB II (%)	-0.3	3.2***	-7.5***
Disability (%)	0.01	0.4	6.1***
Employed (%)	0.3	1.0	-1.0
Income (30 days)	1.7	-71.4***	402.4***

* < 0.05; ** 0.001<p<0.01; *** p < 0.001

Study 1: Main Findings

- Non-consent bias present for some variables
- Overall non-consent biases are small
- NR/ME biases tend to be larger than non-consent biases
  - data linkage makes sense from TSE perspective

Study 1: Limitations

- PASS response rate is low (26.7%)
- Special population (German benefit recipients)
  - Correlates of consent similar in general population
- Quality of administrative data is unknown
- Admin data come from various sources

Study 2: Diabetes Validation Project

- 2006 Health and Retirement Study
  - Longitudinal study of Americans age 50 and older
  - Study began in 1992; biennial interviews
  - Half of Rs randomized to Enhanced Face-to-Face IW
- Medicare administrative claims data
  - 86% consent rate
- Biomarker collection (blood and saliva)
  - 83% consent rate
- Data sources linked for Medicare beneficiaries age 65 and older (N=2,030)

Diabetes Measures

- HRS self-reports
  - "Has a doctor ever told you that you have diabetes or high blood sugar?"
- Medicare claims
  - Chronic Conditions Warehouse algorithm
  - At least one inpatient or two outpatient visits with indication of diabetes (Buccaneer, 2009)
- Blood data
  - Hemoglobin A1c level > 6.5 (clinical threshold)

Validated Diabetes Status

- Combination of self-reports, claims, and blood data
- Definition:
  - Agreement between self-report and claims data
  - At least one diabetes indication and HA1c > 6.5
- Validated diabetes rate (weighted) = 20.4%

Percent Distribution of Diabetics

	Self-Reports	Medicare Claims	Validated
Overall	**20.4**	**27.0**	20.4
Age			
65-74	**53.8**	**45.6**	51.7
75-84	**37.1**	**42.3**	38.8
85+	**9.1**	**12.1**	9.6
N	441	569	441

- No significant differences found for gender, Hispanic ethnicity, race, self-reported health rating, and moderate activity.

Percentage of Correct Diabetes Indications

	N	Self-Reports	Claims
Overall	2030	**94.8**	**73.7**
Age			
65-74	1130	**91.4**	**81.5**
75-84	684	**98.9**	**69.1**
85+	216	**98.3**	**59.8**
Gender			
Male	1187	**95.3**	**77.2**
Female	843	**94.4**	**70.9**
Race			
White	1785	**94.2**	**74.1**
Non-White	245	**100.0**	**70.0**

Discordant Cases by Lab Results

	Self-Report Only	Claims Only	Concordants
Hemoglobin A1c (mean)	**6.32**	**5.86**	6.60
Ha1c > 6.5 (%)	**30.5**	**12.4**	45.6
N	34	162	407

- Claims only cases tend to be older and report better health than concordant cases
- No difference on memory rating or # of diagnoses

Health Care Utilization Outcomes (2006)

Diabetics	Self-Reports	Claims	Validated Standard
Avg. Medicare Reimbursement ($)	9412	9730	9706
Avg. # of Office Visits	9.8	10.4	10.0
Avg # of Hospitalizations	0.4	0.4	0.4
Total # of diabetics	441	569	441

- Utilization unaffected by diabetes definition

Study 2: Main Findings

- Administrative claims tend to overestimate diabetes status compared to self-reports and validated measure.
- Claims-only diabetics tend to be healthier and have lower Ha1c levels compared to SR-only and concordant diabetics.
- Health care utilization outcomes unaffected by either diabetes definition.

Study 2: Limitations

- Validated diabetes measure is imperfect
  - No access to medical records
- Relatively small sample size
- Non-random consent to biomarkers/linkage
- Biomarker collection at a single point in time
  - self-report covers *ever* told
  - prediabetics may have made successful lifestyle changes

Overall Conclusions

- Non-consent biases exist in survey *and* administrative estimates
  - Reassuring: biases are small relative to other errors
- Administrative estimates may conflict with survey estimates
  - Assumption that administrative data is 'gold standard' may not be valid
  - Reassuring: substantive results may be unaffected

To Link or Not to Link?

- It depends…
  - What is being linked?
  - What is the quality of the admin data?
  - What are the researchers objectives?
  - Could data users potentially misuse the linked data, or make invalid inferences?
  - How willing are respondents to consent to linkage?
  - How willing are data agencies to share/release administrative data for linkage purposes?

Future Research

- Assessment of administrative data quality
  - Quality indicators
  - Replication
- Mechanisms of consent
  - Why are some Rs reluctant to consent? How to surmount this problem?
  - Are consent rates correlated with biases?
- Data linkage techniques
  - Statistical matching vs. exact matching

Thank you!

joesaks@umich.edu

Extra Slides

Consent Propensity Model

- Random-effects logistic regression
  - Respondents nested (non-randomly) within interviewers
- Outcome: linkage consent
- Covariates: survey variables
  - socio-demographics
  - paradata (call attempts, panel cooperation)
  - interviewer characteristics (age, education, gender)

Model Summary

- Sociodemographics
  - Age (-), Employed (+)
- Paradata
  - Panel cooperation (+)
- Interviewer characteristics
  - Gender (+), Education (-)
  - Interviewer variance component ($p < 0.05$)
- Model Diagnostics
  - Pseudo $R^2 = 0.05$
  - Adj. Pseudo $R^2 = 0.03$