SURV622/SURVMETH622: Record Linkage Applications and Methods Part 2

February 10, 2025

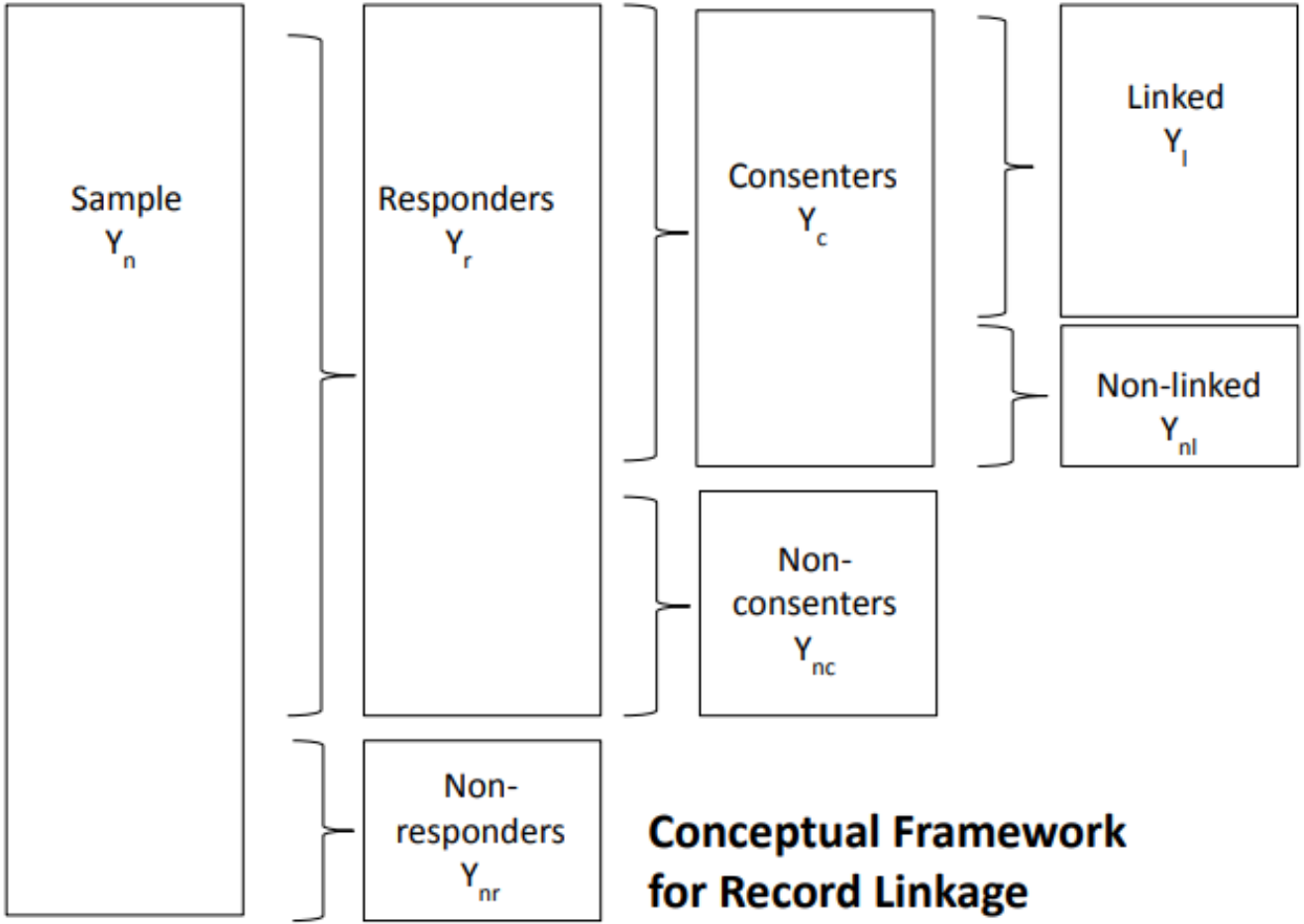
James Wagner

Overview of Potential for Bias in Record Linkage

Conceptual Framework for Record Linkage

- Consider application of record linkage to augment information on a survey data file
- Sample observations may be lost due to:
  - Survey non-response
  - Respondents who do not consent to linkage
  - Consenters for whom no records can be linked
- Similar to nonresponse bias, consent bias (linkage bias) depends not only on share consenting (share linked) but also on differences between those consenting (linked) and those not

$$B(\bar{y}_{YES}) = \bar{y}_{YES} - \bar{y}_t = \left(\frac{n_{NO}}{n}\right)(\bar{y}_{YES} - \bar{y}_{NO})$$



Conceptual Framework for Record Linkage

Source: Sakshaug and Antoni (2013)

Consent Bias in Record Linkage

Factors Affecting Consent Rates

- Growing body of research on factors affecting agreement to record linkage, but difficult to draw general conclusions
- **Demographic variables** often correlated with probability of consent, but direction of effect varies across studies
  - Pattern by **age** and **sex** mixed; **education** more generally associated positively with consent
  - Consent may depend not only on individual characteristics but also on **type of record** for which linkage sought and **interaction between the two**
- Survey item nonresponse and other indicators of resistance to survey participation associated negatively with probability of consent

Evaluating Consent Bias

- To quantify magnitude of consent bias, must be able to compare values of administrative variables for consenters and non-consenters
  - Very few studies have done this
- Sakshaug and Kreuter (2012) examine consent bias in German Labour Market and Social Security (PASS) survey conducted by German Institute for Employment Research (IAB)
  - Sample of 23,736 households receiving means-tested benefits at time sample was drawn
  - Linkage consent rate close to 80%
  - Unique identifiers and access to administrative records available for all sample members
    - Did not link individual survey responses for nonconsenters to administrative records, but used survey paradata to identify respondents/nonrespondents and consenters/nonconsenters
    - Constructed separate estimates based on administrative records for full sample, respondents and consenters

Evaluating Consent Bias

$$(1)\quad Non-response\ bias\,(\bar{y}_{ADMIN}) = \bar{y}_{ADMIN,\mathrm{Resps}} - \bar{y}_{ADMIN,\mathrm{Sample}}$$

$$(2)\quad Non-consent\ bias\,(\bar{y}_{ADMIN}) = \bar{y}_{ADMIN,Consent} - \bar{y}_{ADMIN,\mathrm{Resps}}$$

$$(3)\quad Measurement\ bias\,(\bar{y}_{ADMIN}) = \bar{y}_{\mathrm{PASS,Resps}} - \bar{y}_{ADMIN,\mathrm{Resps}}$$

Source: Sakshaug and Kreuter (2012)

Evaluating Consent Bias

	$Y_{ADMIN, Sample}$	$Y_{ADMIN, Resps}$	$Y_{ADMIN, Consents}$	$Y_{SURV, Resps}$
Age (years)	39.5	39.5	39.3	39.5
Foreign (pct)	16.5	11.0	10.0	8.5
UI benefits (pct)	80.2	83.4	83.1	75.9
Disability (pct)	4.9	5.3	5.3	11.3
Employed (pct)	29.3	30.3	30.6	29.3
Income (Euros)	799.7	728.5	730.2	1130.9

Source: Sakshaug and Kreuter (2012)

Thinking about the biases in percent foreign and using the formulas on the previous slide:

- How would you calculate *nonresponse bias?*
- How would you calculate *nonconsent bias?*
- How would you calculate *measurement error?*

Evaluating Consent Bias

	Nonresponse Bias	Nonconsent Bias	Measurement Error
Age (years)	0.1	-0.3*	-0.0
Foreign (pct)	-5.6*,nc	-0.9*,nr,m	-2.5*,nc
UI benefits (pct)	3.2*,nc	-0.3nr,m	-7.5*,nc
Disability (pct)	0.4nc	0.1nr,m	6.1*,nc
Employed (pct)	1.0	0.3	-1.0
Income (Euros)	-71.4*,nc	1.7nr,m	402.4*,nc

* significantly different from zero, nc significant different from nonconsent bias, nr significantly different from nonresponse bias, m significant different from measurement error, all at 0.05 level

Source: Sakshaug and Kreuter (2012)

Optimizing Consent Rates

Features of Consent Request May Affect Agreement

- Consent rates for linkage vary widely across surveys
  - Some evidence of declining consent rates for US surveys over time (Fulton 2012)
- Even if non-consent does not cause bias, having fewer observations increases variance
- Research on optimizing consent rates
  - **Placement** of consent request
  - **Framing** of consent request
  - **Opt in** (active) versus **opt out** (passive) consent

Placement of Consent Request

- Conventional to place record linkage requests at end of survey interview
  - Intuition: Allows time for interviewers to develop rapport and trust with respondent, nonconsent does not interfere with response to survey
  - Research suggests this is not the best option
- Available evidence suggests either of two alternatives produces higher consent rates
  - Option 1: Place request at start of interview
  - Option 2: Place request in section of survey asking questions about topic to which records pertain (i.e., in context)

Placement and Consent Rates

- Saukshaug, Tutz, and Kreuter (2013) report on experiment in telephone survey of German residents with various employment histories done for the IAB
  - Respondents told "We would like to use for the analysis of the survey data parts of the data which are available at the [IAB]… This is for example additional information of previous periods of employment, unemployment, and participation in active labor market programs during unemployment. In order to merge this data to the interview data I would cordially ask you to agree. Do you agree with it?" (English translation from German)
  - No refusal conversion efforts
  - Agreement significantly higher when request placed at beginning of interview (95.6%) than at end of interview (86.0%)

Placement of Consent Request

- Sala, Knies, and Burton (2014) report on experiment in computer-assisted face-to-face (CAPI) interviews in longitudinal panel survey of UK residents
  - Experiment conducted in wave 4 of study
  - Respondents asked for consent to link to administrative records on benefit receipt and, if they agree, to sign a consent form
    - Investigators do not plan to actually link to administrative records
  - Agreement significantly higher when request made immediately following benefit module in the survey (65%) than at end of the survey (58%)

Framing of Consent Request

- Common in record linkage requests to emphasize the **benefits** of consenting
  - Greater accuracy: "The best place to get this information is [source]. We are therefore asking…"
  - Reduced burden for survey respondent: "To keep the interview as short as possible, we would like…"
- Several studies have found no effect of mentioning benefits on consent rates in interview-administered surveys, but one study found that mention of **time savings** in a web survey had a positive effect (Sakshaug and Kreuter 2014)
  - Possible that mention of benefits more salient in visually-centric modes

Framing of Consent Requests

- Recent research has tested whether there are differences in behavior when agreement is framed as producing a benefit (**gain framing**) versus avoiding a loss (**loss framing**)
  - Rests on foundational research on prospect theory by Kahnemann and Tversky (1979, 1984, 1992)
- Previous research by Tourangeau and Ye (2009) tested idea in context of soliciting a second interview from people who had completed an earlier interview
  - Gain framing: "The information you've already provided to use will be a lot more valuable if you complete the second interview."
  - Loss framing: "Unfortunately, the information you've already provided to use will be less valuable unless you complete the second interview."
  - Agreement rates significantly higher with **loss** framing (87.5%) than with gain framing (77.9%)

Framing of Consent Request

- Kreuter et al (2016) apply same idea to record linkage requests
  - Telephone survey of Maryland residents asked for consent to link to voting records
  - **Gain** framing: "The information you have provided so far would be a lot more valuable to use if we could link to public voting records. Do we have your permission to link your answers to your voting record?"
  - **Loss** framing: "The information you have provided so far would be much less valuable to use if we can't link it to public voting records. Do we have your permission to link your answers to your voting record?
  - Consent rate significantly higher with **loss** framing (66.8%) than with gain framing (56.1%).

Opt-in (Active) vs. Opt-out (Passive) Consent

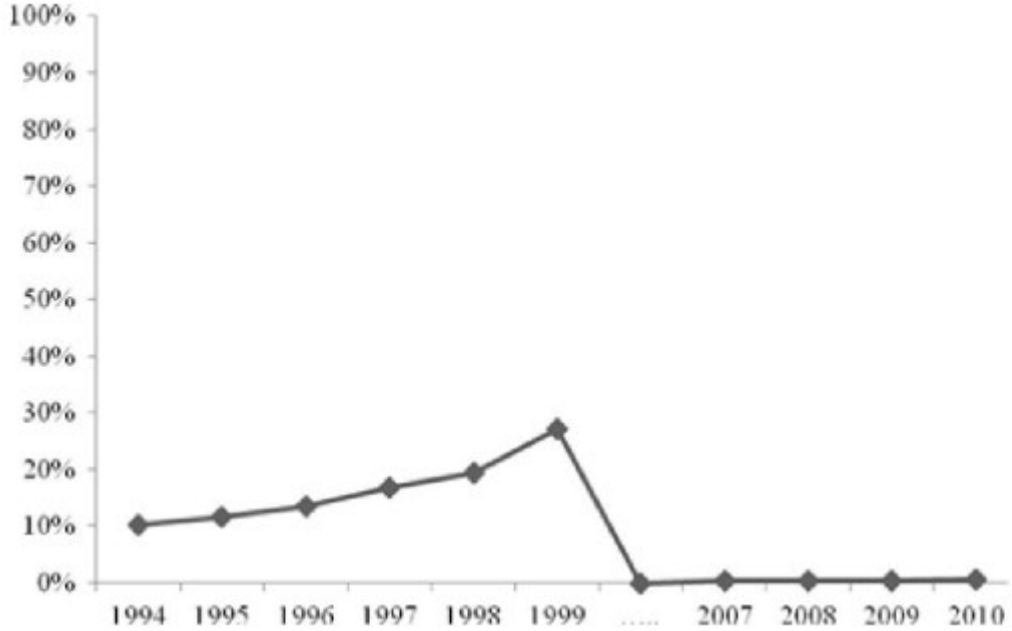
- Active consent requires respondents to provide explicit agreement for linkage
  - May sign a form, provide a linkage ID, or provide oral confirmation
- Passive consent requires respondents to object if they do now want linkage to occur
- Consent to linkage higher with **opt-out** (passive) procedures than with opt-in (active) procedures
  - Consistent with a large body of research in other contexts showing that the default option often matters a lot

Opt-in (Active) vs. Opt-out (Passive) Consent

- Example: Census Bureau procedures for assigning Personal Identification Keys (PIKs) to CPS records for record linkage
  - Prior to 2006, respondents asked to opt in and provide SSN for linking
  - Since 2006, consent has been opt-out and SSN not requested (use other information for linking)

> Linkage in the post-2006 pre-notification letter: "Occasionally, we may combine data from the CPS with data we obtain from other government agencies to provide a comprehensive set of summary information about employment, income, and participation in various government programs. The same confidentiality laws that protect your survey answers also protect any additional information we collect (Title 13, United States Code, Section 9.) To ensure your protection, the laptops used for the data collection are password protected and all survey responses are encrypted. *If you wish the request that your information not be combined with information we obtain from other agencies, we ask that you notify the Field Representative at the time of the interview* [emphasis added]." (U.S. Census Bureau, 2008)

Rates of Consent Refusal to Link Survey and Administrative Data in the Current Population Survey



Source: Fulton (2012)

Some Questions About Record Linkage Consent Procedures

- Do respondents understand what is involved in record linkage?
  - Some evidence to suggest that many respondents do not understand clearly what it is they are agreeing to
  - Providing more details can improve understanding
    - Analyzing data form an experiment on the LISS, Das and Couper (2014) find that respondents who receive more detailed explanations of linkage scored higher on later questions about linkage procedures
    - In this study, more detailed explanations were not associated with lower opt-out consent rates
- Are there reasons to be concerned about the potential harm to respondents from record linkage?
  - Potential harms in the context of producing statistical information arguably are minimal
  - Related to protections that are in place to protect respondents' privacy and confidentiality

Privacy, Confidentiality, and Data Linkage

Data Linkage Considerations

- Suppose we go through all of the steps of linking two datasets together (e.g., one administrative, one survey). We go through extensive pre-processing, use blocking, use probabilistic linkages, and choose a threshold that give the best accuracy/precision/metric of choice. Suppose everyone gave consent, and we were able to match a high proportion of people.
  - Does being able to match a high proportion of people give you confidence? Where might errors come in? How could this lead to bias?
  - How might this affect our analysis?

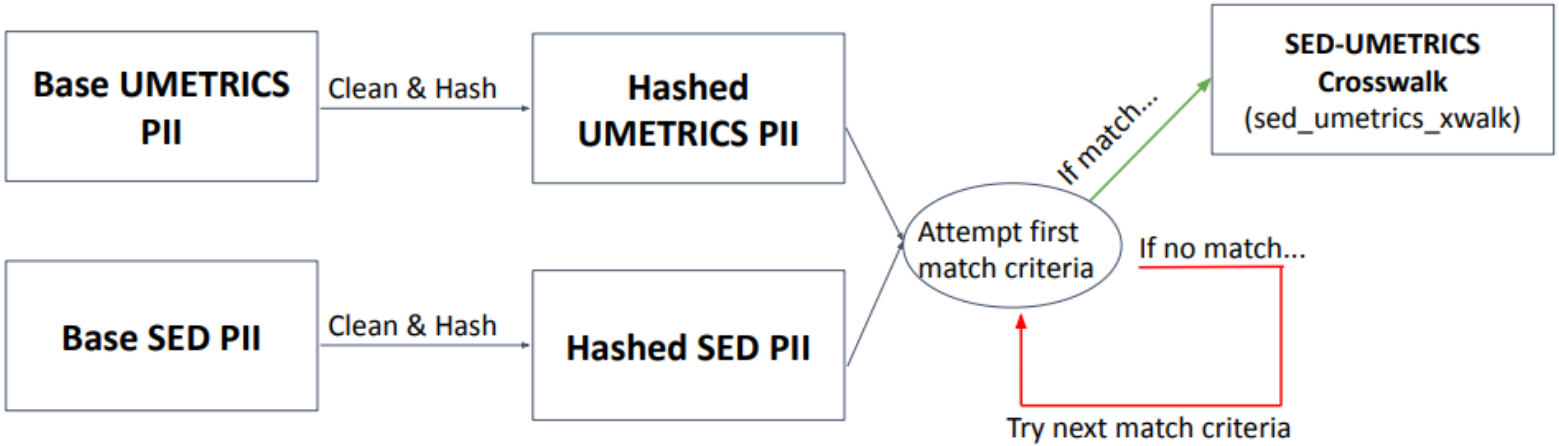
Example: NCSES Data

- Recall: The Survey of Earned Doctorates/Survey of Doctorate Recipients (SED/SDR) has information about PhDs and their employment. UMETRICS has information about grant funding.
- To link these, names were pre-processed (nickname dictionary, lowercase, etc.; "name disambiguation") then hashed by data owners to protect privacy, then shared to link on hashed names.
  - What are the sources of bias here?
  - How might this affect our analysis?
  - We blocked on institution. What are some issues that might arise?

Example: NCSES Data

- Due to privacy concerns, we don't have the luxury of doing the iterative matching and pre-processing
- We need to make some decisions on how to approach the matching process.
- Bias from matching on names:
  - Non-English names might have higher chance of failing to match because of incomplete nickname dictionary, failure to account for structure of name, etc.
    - E.g., Ekaterina/Katya, Natasha, Kim Jungkyun/Jung-Kyun Kim
  - This might mean we have more complete data on domestic students.
    - If we conclude that foreign students have a harder time finding funding, is that really the cases, or is it just a case of poor matching?

Linking SED and UMETRICS Individuals



Match Criteria and Order:

1. University + Given (First Name + Middle Name) + Family (Last Name) + Birth Year + Birth Month
2. University + First Word of Given + Last Word of Given + Family + Birth Year + Birth Month
3. University + First Word of Given + Family + Birth Year + Birth Month
4. University + Given + Family + Birth Year
5. University + First Word of Given + Last Word of Given + Family + Birth Year
6. University + First Word of Given + Family + Birth Year

Addressing Bias

- Possible steps to take:
  - Consider how the pre-processing is done. Can we address issues before they arise?
  - Check accuracy/precision/metric of choice for different subgroups. What is the match rate?
  - Check match rate according to blocking variable. Does the match rate differ significantly? If so, why? You don't expect them all to e the same, but make sure there isn't a systematic issue.
- Note: We are not necessarily checking for universally good match rates.