# Ethical Issues in Linking Non-Survey Data to Survey Data: A Practical Guide

## INTRODUCTION

**This guide identifies ethical issues and possible solutions for researchers who are considering linking non-survey data to survey data. Our focus is on record-level linking of one person's survey interview to non-survey data pertaining to that person, but we discuss some ways that aggregating to other levels of data can help manage some ethical challenges.**

### FINE PRINT

Many legal requirements pertain to data linkage and these requirements change frequently, much more so than many other aspects of survey research. These requirements may be related to who can access non-survey data sources, data privacy laws and rules about data sovereignty, use of linkage information or personally identifying information, or about data storage and archiving.

This guide addresses ethical issues when there are no legal requirements, or when researchers may want to take precautions beyond what is legally required. We urge researchers using this guide to investigate any laws that may apply to the linkage that you are considering, and also to be especially mindful about whether or not your research could pose risks to vulnerable populations.

### FOOTNOTE

There are other issues that are important to consider when linking non-survey and survey data – such as data quality of linked files or logistical challenges to linking that don't have ethical implications, or statistical or technical issues associated with creating and working with linked files - but this guide doesn't address those.

## UNDERSTANDING NON-SURVEY DATA

### WHAT ARE NON-SURVEY DATA?

Non-survey data generally start outside of a research context. These data can come from a government program, a human service like schooling or health care, or from a commercial activity like a transaction or manufacturing. People also generate non-survey data when they use technology such as cell phone location trackers or social media accounts.

### WHAT'S SPECIAL ABOUT NON-SURVEY DATA FOR RESPONDENTS?

Respondents may not understand:

- that a data source exists
- that a researcher could track down a person's data
- that data can be linked to their survey responses
- how data could be linked to their survey responses
- who can access data about them, how data are accessed, and for how long data can be accessed
- what can be learned about individuals by linking non-survey data to survey responses
- what impact linked data could have on an individual, especially if disclosure occurs

All of these factors may change over time for non-survey data more and faster than they would for survey data. For example, cell-phone location data, facial recognition data from public security cameras, or a car insurance company's ability to know how often drivers speed or brake abruptly would all have been inconceivable just a few years ago.

AAPOR
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

## KEY QUESTIONS FOR IDENTIFYING ETHICAL CHALLENGES FOR A SURVEY DATA/NON-SURVEY DATA LINKAGE

As researchers, we have ethical obligations to respondents in how we handle their data. We also have obligations to our fellow researchers and to the audiences who would learn from our research. Here are some questions researchers can ask themselves when planning linkages:

- *Can respondents be asked to provide consent?* Current contact information, financial resources, and elapsed time are all required to gain consent and can be particularly challenging if the survey took place a long time ago.
- *What did the consent language mean to the respondents?* Especially if consent for linkage was collected many years ago, respondents may have reasonably understood the terms and requests differently from how they would be interpreted today. Responsible researchers interpret consent language as respondents would have done at the time of consent.
- *How accurate will the linkages be?* If linking variables are unreliable, or the data to be linked are not clean or suffer from quality issues, the linked data may be flawed. What are the implications for respondents if the linked data are flawed: could they be re-identified or suffer harm? Could flaws in the linked data lead to misleading analyses?
- *Who has access to the survey data, the linking variables, or the non-survey data to be linked?* Does limited data access raise questions about the ethics of linking? Can re-analysis or replication occur given the limited data access?
- *Could describing the linkage strategy and linked files increase disclosure risk? How we report linkage strategy can be as important to limiting disclosure as the reported estimates themselves. Researchers can be thoughtful about reporting details that increase risk to participants without improving transparency.*

## TOOLS YOU CAN USE: WAYS TO SUCCESSFULLY MANAGE ETHICAL ISSUES

- Gaining informed consent from respondents to conduct a linkage is a best practice where possible. Many fewer concerns arise when the participants have consented to a specific linkage after understanding the data, how they will be linked, and why they will be linked.
- Useful insights can sometimes be gained without record-level linkage. If consent is not available, linking variables are limited, or disclosure risk is high, it may be possible to choose a coarser level of linkage between the two data sources. For example, workers' survey data can be linked to community-level average income data when it isn't possible to link individual workers to their own tax records.
- How analyses are reported can increase or decrease the chance of harms to respondents. Detailed counts of individuals in small subgroups can give away a lot, while regression coefficients about patterns of behavior can provide research insights with very little risk of harm to respondents.
- Carefully assessing disclosure risk and data quality limitations of linked data lets researchers be sure they can minimize harms to respondents.
- We can document our source data, linkage strategies, and linked data to help our audiences correctly interpret insights from the linkage. Patterns of non-response, missingness, refusal to consent to linkage, or linkage match rates can be especially informative but invisible from most analyses or linked data files unless we make it a point to report them.
- Protect the linked data and provide data access to minimize potential harms to respondents but maximize opportunities for equity in future research.
- Develop policies for communicating with respondents: is there anything to proactively communicate to them, such as data breaches or actionable health or legal risks revealed by the linked data? What information will be available if a respondent contacts the researcher with questions? Are any remedies available, such as data removal on request?

# VIGNETTES

Below are three vignettes intended to illustrate examples of the use of data linkages best practices in different research scenarios.

## RECONTACTING RESPONDENTS TO GET LINKAGE CONSENT

Trina Patel and her team wish to conduct a study that examines political attitudes by linking survey responses to publicly available social media posts. They would assign users a partisanship score based on their posts and then link this score to survey data that includes information on income levels, occupation, and industry. While the social media data are technically public, Patel and her team recognize that respondents may not have anticipated such linkages when they participated in the survey. In addition, the research team realizes that there may be some disclosure risk to individuals when social media posts are paired with less common jobs. While they are exploring options for de-identifying data and assessing disclosure risk, a member of the research team realizes that the team could try to recontact survey respondents through their social media accounts. Using private outreach to respondents through their social media accounts, the team is able to gain very high rates of respondent consent to link. In addition, the team devises an analytic strategy that reports partisanship data for broad employment categories, ensuring that disclosure risk to respondents is minimized.

## INTERPRETING PRIOR CONSENT LANGUAGE IN NEW CONTEXTS

Dr. Jacobs and his team conducted a survey in 1990 where they secured consent to link survey data to health insurance records. They would now like to link the survey data to electronic health records that would allow analyses of long-term outcomes. None of the original consent language addressed the breadth of content available in electronic health records. Recontact of respondents to gain new consent was not feasible. Jacobs and his team worked with their institutional review board, the data access team for the electronic health records company, and the legal team for the health insurance company originally involved to assess the ethical and legal issues. They jointly reviewed the original consent language, legal developments since then, and customary practices today. They determined that a subset of electronic health records information was consistent with the original consent language and with prevailing standards. The team successfully analyzed this subset of data in conjunction with the survey data to gain valuable insights on long-term health outcomes related to the original survey design.

## REASONABLE AWARENESS

A survey was conducted of workers in a union. Workers were informed that they had been sampled from union files, and the questionnaire content referred to specific aspects of job conditions in the union files. Researchers would like to re-analyze these data pulling in additional information from the union files. No consent to link was originally secured and current contact information is not available. Union officials, associated employers and the researchers' ethics committee agreed that since respondents were originally aware that union files were used in the survey administration, and since the content was not determined to be sensitive, it was agreed that the proposed analysis could proceed.

*Prepared by AAPOR ad-hoc group on ethics of non-survey linkages:*
*Josiane Bechara, Trent D. Buskirk, A. Rupa Datta, Ned English, and an additional colleague, 2025.*