SURVEYS · TECHNOLOGY AND THE FUTURE

Administrative Data Linkage with Surveys: Expanding the Scope of Generated Knowledge

Linking survey answers to administrative records, such as tax, benefit, health and education records, can improve accuracy, reduce burden and answer questions neither source can answer alone. This article explains how linkage works, its benefits, the problems of consent, linkage error and privacy, and how to carry it out well.

ModQu research article · 2,612 words · 24 references (APA 7th edition)

Governments and organisations hold vast amounts of information about people in **administrative records**: tax returns, benefit payments, hospital admissions, school results, employment histories and many others. These records are created to run services, not for research, but they contain detailed and often accurate information about many aspects of people's lives. Surveys, by contrast, are designed for research and can measure attitudes, experiences and circumstances that no administrative system records.

Data linkage connects the two. With respondents' permission, or under legal authority, survey answers are matched to administrative records for the same people. The linked data can then be used to check the accuracy of survey answers, to add information that would be burdensome or impossible to ask about, and to follow people over time without further interviews. Linkage has become one of the main ways in which surveys are being combined with other data sources (Groves, 2011; Connelly et al., 2016). This article explains how it works, what it offers and the problems that must be managed.

A short history

Record linkage has a long history in health research and official statistics. In the mid-twentieth century, researchers began linking birth, marriage, hospital and death records to study families and diseases over time. As computers made it possible to compare millions of records, the problem of matching records that did not share a reliable identifier became central. Fellegi and Sunter (1969), working at Statistics Canada, provided a formal statistical theory that showed how to weigh agreements and disagreements between records and how to set thresholds that balance false matches against missed matches.

Linking surveys to records followed. Validation studies in the 1970s and 1980s compared survey answers about income, hospital stays and other events with records to measure reporting error (Bound et al., 2001). Over time, linkage moved from a method for checking surveys to a way of enriching them, and many large surveys now ask all respondents for permission to link. In health research, large linked data systems that connect hospital, primary care and population records have become standard tools, and their experience has shaped practice in social surveys (Bohensky et al., 2010).

What administrative data offer and lack

Administrative records have several strengths. They often cover the whole population that uses a service, not a sample. They record events as they happen, without relying on memory. They can be very precise, for example on income received or dates of hospital stays. And they already exist, so using them adds little cost or burden to respondents.

They also have limits. They cover only people who come into contact with a particular system, and they record only what the system needs. Definitions are set for administrative purposes and may not match research concepts: a tax system's definition of income or a benefit system's definition of a household may differ from those used in research. Records can contain errors, be incomplete or change when rules change (Hand, 2018). And they say nothing about people's attitudes, intentions or reasons.

Surveys and administrative data therefore complement each other. Figure 1 summarises what each contributes.



Figure 1. What surveys and administrative records each contribute. Diagram by ModQu.

How linkage works

Deterministic and probabilistic linkage

The simplest form of linkage uses a unique identifier, such as a national identity or social insurance number, recorded in both sources. Records that share the identifier are matched. This is called **deterministic** linkage.

Often no such identifier is available, or it contains errors. Records must then be matched on combinations of names, dates of birth, sex and addresses, which may be spelled differently or have changed. **Probabilistic linkage** calculates, for each possible pair of records, how likely it is that they belong to the same person, based on how many characteristics agree and how informative each agreement is. Fellegi and Sunter (1969) set out the statistical theory for this approach, and it remains the basis of most modern linkage methods (Herzog et al., 2007; Christen, 2012).

Privacy-preserving linkage

To protect privacy, linkage is often carried out by a trusted third party, so that researchers never see names and addresses. The data holders send identifying information to the linkage unit, which creates a key linking the records; researchers then receive only the linked research variables with the key, not the identifiers. Some methods encrypt identifiers before linkage, so that even the linkage unit does not see them in plain form.

Consent and legal basis

Survey linkage usually requires either respondents' consent or a specific legal basis. Many surveys ask respondents during the interview whether they agree to their answers being linked to particular records. In some countries, statistical agencies have legal powers to link data without consent for statistical purposes, subject to strict safeguards. Data protection law sets conditions for both approaches (European Parliament & Council of the European Union, 2016).

Uses of linked data

Checking and improving survey measurement

Comparing survey answers with records shows how accurate the answers are. Such validation studies have found that people often under-report income, especially from benefits and investments, and misreport dates and durations of events (Bound et al., 2001; Moore et al., 2000). Meyer et al. (2015) showed that under-reporting of government benefits in major American household surveys had increased over time. Using linked data, Meyer and Mittag (2019) found that correcting survey income with administrative records substantially changed estimates of poverty and of the effect of government programmes.

Kreuter et al. (2010) used linked employment records to study both nonresponse and measurement error in a German labour market survey. They could compare respondents with non-respondents using the records, and compare respondents' answers with what the records showed. Both kinds of error affected estimates, and the larger source of error varied between variables.

Reducing burden

If reliable information is available in records, there is no need to ask for it. Some surveys replace detailed questions on income, benefits or health service use with linked records, shortening the interview and reducing the risk of reporting errors. Records can also be used to prompt respondents rather than replace their answers. In **dependent interviewing**, information already known about a respondent, from an earlier interview or from records, is fed back into the questions, for example by reminding respondents of income sources they reported before and asking whether they still receive them. Lynn et al. (2006) found that this approach reduced the under-reporting of income sources in a British panel survey. Using record information in this way requires the same consent and protection as other forms of linkage, because respondents see that the survey already holds information about them, and some may find this unwelcome if it is not explained.

Extending the scope of surveys

Linked records can follow respondents beyond the survey, for example recording later hospital admissions, earnings or deaths. This turns a single survey into a long-term study at little extra cost. Conversely, surveys add context to administrative data: why people used a service, what they thought of it and what else was happening in their lives.

Studying nonresponse

When a survey sample is drawn from an administrative register, information in the register is available for both respondents and non-respondents. This allows a direct check of how they differ and how much nonresponse biases estimates, which is rarely possible otherwise (Kreuter et al., 2010).

The problems

Consent and consent bias

Not all respondents agree to linkage. Consent rates vary widely between surveys, types of record and countries. Jenkins et al. (2006) studied consent in a large British household survey and found that consent differed by respondent characteristics and was related to the type of record requested. Sala et al. (2012) found that interviewer characteristics and the way the request was made also affected consent. Sakshaug et al. (2012) studied the mechanisms behind consent decisions and found that respondents' trust, concerns about privacy and engagement with the survey all played a role.

If those who consent differ from those who do not, the linked data are biased. Sakshaug and Kreuter (2012) used administrative data available for a whole sample to measure this bias directly. They found that non-consent bias existed for some variables but was often smaller than bias from nonresponse or measurement error. The size of consent bias should be assessed in each study, not assumed. Singer (2003) showed that people's decisions about consent depend on how they perceive the risks and benefits, which researchers can address by explaining clearly how linked data will be used and protected.

Figure 2 illustrates a typical pattern of consent rates for different types of record.

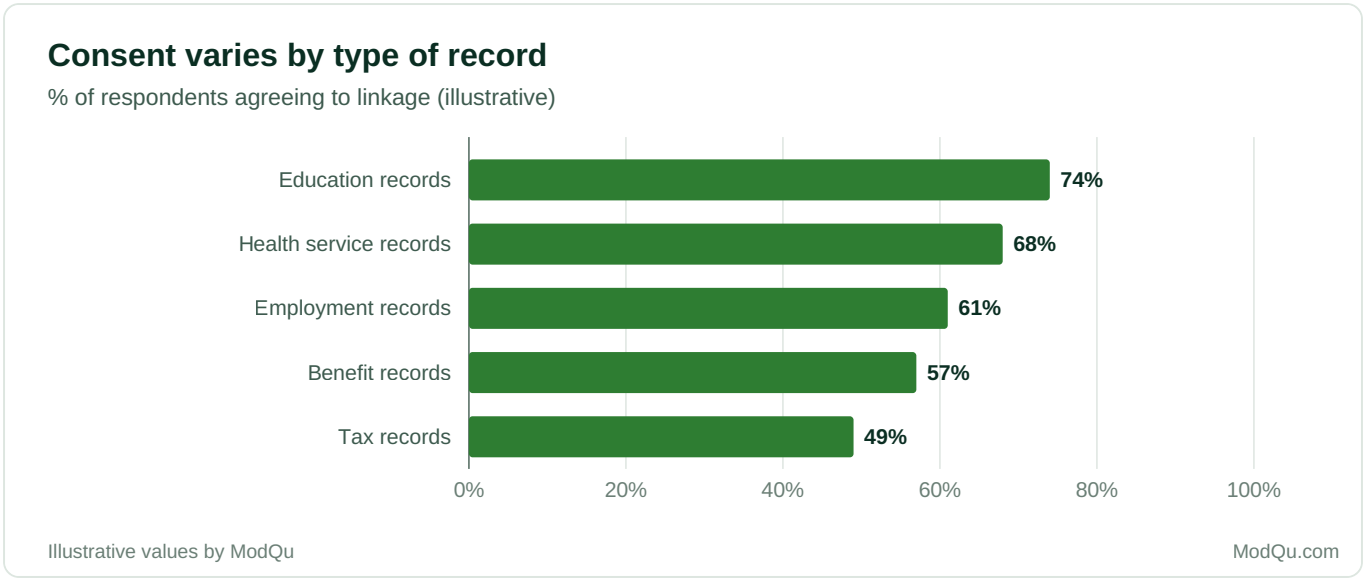


Figure 2. Consent to linkage by type of record in a hypothetical survey (illustrative values). Chart by ModQu.

Linkage error

Linkage is rarely perfect. **Missed matches** occur when records belonging to the same person are not linked, for example because of spelling differences or changes of address. **False matches** occur when records belonging to different people are linked. Both kinds of error can bias results, especially if they are more common for some groups, such as people with names that are transcribed in different ways, people who move often or people with incomplete records (Bohensky et al., 2010). Doidge and Harron (2019) describe how linkage errors can bias estimates and how to assess and correct for them. Linkage quality should be measured and reported, and analysts should check whether linkage rates differ between groups.

Differences in definitions and timing

Survey questions and administrative records may measure different things. A survey may ask about household income in the last year, while tax records report individual income for a tax year. Benefit records may count payments to a household differently from the way a survey defines a household. Such differences are not errors in either source, but they must be understood before records are used to validate or replace survey answers (Hand, 2018).

Privacy and disclosure risk

Linked data sets are detailed and can be more identifying than either source alone. Combining a few characteristics from a survey with precise dates and amounts from records may make individuals recognisable, even without names (Sweeney, 2002). Linked data are therefore usually made available only in secure settings. The **Five Safes** framework sets out five questions for managing access: whether the project is appropriate, whether the people are trusted, whether the setting is secure, whether the data are suitably protected and whether the outputs are safe to release (Desai et al., 2016).

Access and practical barriers

Gaining access to administrative data can take a long time. Data holders may lack the legal basis, resources or willingness to share. Linked data may arrive years after a survey was carried out. Harron et al. (2017) describe these practical barriers and argue for clearer legal frameworks, better documentation and more transparency about linkage methods.

Analysing linked data

Analysts of linked data need to account for the processes that produced it. Three steps are especially important.

First, **define the analysis population**. Linked analyses usually cover only respondents who consented and whose records were found. Analysts should state this clearly and compare the linked group with the full sample.

Second, **adjust for selection**. If consent or successful linkage depends on characteristics measured in the survey, weights can be adjusted so that the linked group resembles the full sample on those characteristics. This works only for differences related to measured characteristics, so sensitivity analyses that test how results change under different assumptions are useful (Doidge & Harron, 2019).

Third, **account for linkage uncertainty**. In probabilistic linkage, some links are more certain than others. Analysts can check whether results change when only the most certain links are used, or use methods that incorporate the probability that each link is correct.

When records are used to replace survey answers, analysts should also consider whether the two sources measure the same concept. Replacing a survey question about household income with individual tax records, for example, changes the meaning of the variable, not just its accuracy.

Linkage beyond official statistics

Linkage is not limited to government data. Companies link customer satisfaction surveys to records of purchases, complaints and account closures to see which experiences predict whether customers stay. Employers link engagement surveys to records of absence and staff turnover. Universities link student surveys to academic records.

The same principles apply. Respondents should be told clearly that their answers will be linked to records and for what purpose, and confidentiality promises must be kept. Linking an apparently anonymous staff survey to personnel records, for example, can make individuals identifiable to their managers and break the trust on which honest answers depend. Where linkage is planned, the survey should be described as confidential rather than anonymous, and results should be reported only for groups large enough to protect individuals.

Good practice

- **Plan linkage from the start.** Decide which records will be linked, on what legal basis and with which identifiers, before the survey begins.

- **Ask for consent clearly.** Explain which records, for what purpose, how data will be protected and how long the permission lasts. Test the wording and train interviewers to present it neutrally.
- **Measure consent bias.** Compare those who consent with those who do not, using survey data or sampling frame information, and report the results.
- **Measure linkage quality.** Report match rates, estimated false match rates and differences in linkage rates between groups.
- **Understand the records.** Learn how each administrative variable is defined and produced before using it.
- **Protect privacy.** Separate identifiers from research data, use secure access arrangements and check outputs before release.
- **Document everything.** Users of linked data need to know how linkage was carried out and its limitations.

The role of linkage in the future of surveys

Statistical agencies increasingly see linkage as central to their work. Citro (2014) argued that official statistics should move from relying on multiple survey modes to combining multiple data sources, with surveys focusing on what only they can measure. Linkage is the practical means of doing this. It also changes the role of surveys: from being the only source of information to being one source among several, providing the measures of attitudes, experiences and context that give meaning to administrative counts.

This shift requires public trust. People are more likely to accept linkage when they understand its purpose, believe their data are protected and see the benefits of the research. Survey organisations that explain linkage openly, protect data carefully and publish the knowledge it produces help maintain that trust.

An example

A national study of young adults asks 8,000 respondents about their education, work and wellbeing. With consent, it links their answers to education and employment records. Of the respondents, 72 per cent consent to the education link and 61 per cent to the employment link. Those who consent are slightly more likely to be in work, so the study weights the linked analyses using survey information on employment status.

The linked data show that respondents under-report periods of unemployment lasting less than three months, and that the effect of early school leaving on later earnings is larger than survey answers alone suggest. The survey adds what the records cannot: respondents' reasons for leaving education and their views of their own prospects. The study reports consent rates, linkage rates and the steps taken to adjust for consent bias (illustrative example).

Conclusion

Linking survey answers with administrative records expands what surveys can tell us. It improves measurement, reduces burden, extends studies over time and allows direct checks on nonresponse and reporting error. It also brings new challenges: consent and consent bias, linkage error, differences in definitions and serious privacy risks. Studies that plan linkage carefully, ask for consent openly, measure and report linkage quality and protect data securely can combine the strengths of both sources and generate knowledge that neither could produce alone.

References

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., & Brand, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Services Research, 10*, Article 346. https://doi.org/10.1186/1472-6963-10-346

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier. https://doi.org/10.1016/S1573-4412(01)05012-7

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology, 40*(2), 137–161.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five Safes: Designing data access for research* (Economics Working Paper Series No. 1601). University of the West of England.

Doidge, J. C., & Harron, K. L. (2019). Reflections on modern methods: Linkage error bias. *International Journal of Epidemiology, 48*(6), 2050–2060. https://doi.org/10.1093/ije/dyz203

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A, 181*(3), 555–605. https://doi.org/10.1111/rssa.12315

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717745678

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer. https://doi.org/10.1007/0-387-69505-2

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A., & Sala, E. (2006). Patterns of consent: Evidence from a general household survey. *Journal of the Royal Statistical Society: Series A, 169*(4), 701–722. https://doi.org/10.1111/j.1467-985X.2006.00417.x

Kreuter, F., Müller, G., & Trappmann, M. (2010). Nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly, 74*(5), 880–906. https://doi.org/10.1093/poq/nfq060

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2006). The effects of dependent interviewing on responses to questions on income sources. *Journal of Official Statistics, 22*(3), 357–384.

Meyer, B. D., & Mittag, N. (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics*, *11*(2), 176–204. https://doi.org/10.1257/app.20170478

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226. https://doi.org/10.1257/jep.29.4.199

Moore, J. C., Stinson, L. L., & Welniak, E. J., Jr. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics, 16*(4), 331–361.

Sakshaug, J. W., & Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*, *6*(2), 113–122. https://doi.org/10.18148/srm/2012.v6i2.5094

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research*, *41*(4), 535–569. https://doi.org/10.1177/0049124112460381

Sala, E., Burton, J., & Knies, G. (2012). Correlates of obtaining informed consent to data linkage: Respondent, interview, and interviewer characteristics. *Sociological Methods & Research, 41*(3), 414–439. https://doi.org/10.1177/0049124112457330

Singer, E. (2003). Exploring the meaning of consent: Participation in research and beliefs about risks and benefits. *Journal of Official Statistics*, *19*(3), 273–285.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648