Künn, Steffen

Article

The challenges of linking survey and administrative data

IZA World of Labor

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Künn, Steffen (2015) : The challenges of linking survey and administrative data, IZA World of Labor, ISSN 2054-9571, Institute for the Study of Labor (IZA), Bonn, https://doi.org/10.15185/izawol.214

This Version is available at:
https://hdl.handle.net/10419/125437

STEFFEN KÜNN
Maastricht University, Netherlands, and IZA, Germany

I Z A
World of Labor
Evidence-based policy making

The challenges of linking survey and administrative data

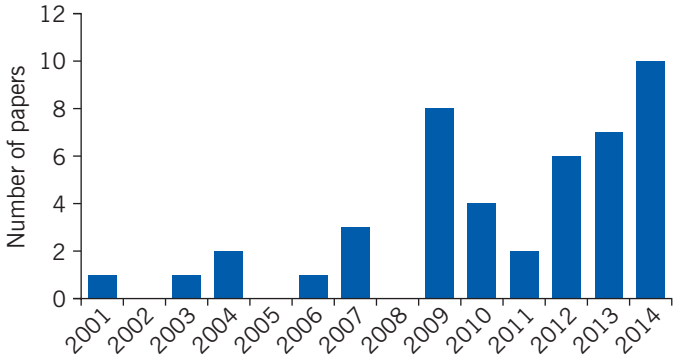
Combining survey and administrative data is growing in popularity, even though data access is still highly restricted

Keywords: record linkage, administrative records, survey data

ELEVATOR PITCH

Using administrative records data and survey data to enhance each other offers huge potential for scientific and policy-related research. Two recent changes have expanded the potential for creating such linked data: the improved availability of data sources and progress in data-matching technology. These developments are reflected, among other ways, in the growing number of academic papers in labor economics that use linked survey and administrative data. While the number of studies using linked data is still small, the trend is clearly upward. Slowing the growth, however, are concerns about data security and privacy, which impede data access.

The use of linked administrative and survey data in labor economics papers has been rising



Source: Calculations based on author's analysis of the IZA discussion paper series.

IZA World of Labor

KEY FINDINGS

Pros

- Data linkage overcomes some of the shortcomings of the two separate data sources.
- Data linkage opens new research opportunities by combining highly reliable administrative data with detailed survey data.
- Administrative records, already collected routinely, are a cheap and authoritative source of data for enriching survey data.
- Data linkage can lower survey costs by requiring fewer questions.
- Data linkage enables sensitive data, such as wages, to be drawn from administrative records, reducing the burden on respondents and likely lessening survey dropout and item nonresponse rates.

Cons

- Linking data can be very costly and time-consuming, mainly because of drawn-out negotiations with data providers.
- Privacy concerns and resulting legal constraints and the need for data anonymization restrict data access and content.
- Requesting consent to use the linked data may introduce consent bias (consenters differ from non-consenters) or may reduce response rates, introducing yet another selection bias.
- Sound linkage requires a unique identifier for each individual; without such identifiers, linkage becomes burdensome and matching quality may suffer.

AUTHOR'S MAIN MESSAGE

Data linkage opens new research opportunities by combining highly reliable administrative records with detailed survey data. Researchers wishing to link the two data sources should establish that both data sources include unique personal identifiers and that the survey includes a properly worded consent request for respondents. Most important, any data security concerns of the data provider (typically a government institution) must be resolved in advance, to avoid having data security concerns lead to restrictions on access or to demands for strict anonymization of the data, which reduce its research potential.

MOTIVATION

The greater availability of survey data and administrative records has had a great impact on the quantity and quality of empirical research in labor economics in recent decades. A relatively new trend is to link both data sources to enhance the research potential of the data. Data linkage is a promising and innovative strategy as it supplements highly reliable administrative records with survey information that is crucial for the statistical analysis but is usually unobserved in the administrative records. As a result, the linked data set contains a large number of variables providing an optimal data source for statistical analysis.

This research potential has been recognized by the scientific community, as illustrated, for instance, by the increasing number of IZA discussion papers using linked data since 2001. However, despite the clear advantages of data linkage, practical implementation remains challenging and time-consuming. Restrictions placed on the use of the data are reflected in the low absolute number of IZA discussion papers using linked data for example. While most of the technical problems in linking survey and administrative data have been solved, the main challenges for research purposes spring from legal constraints related to privacy concerns. These constraints have led to restrictions on data use, as well as consent bias and limited access.

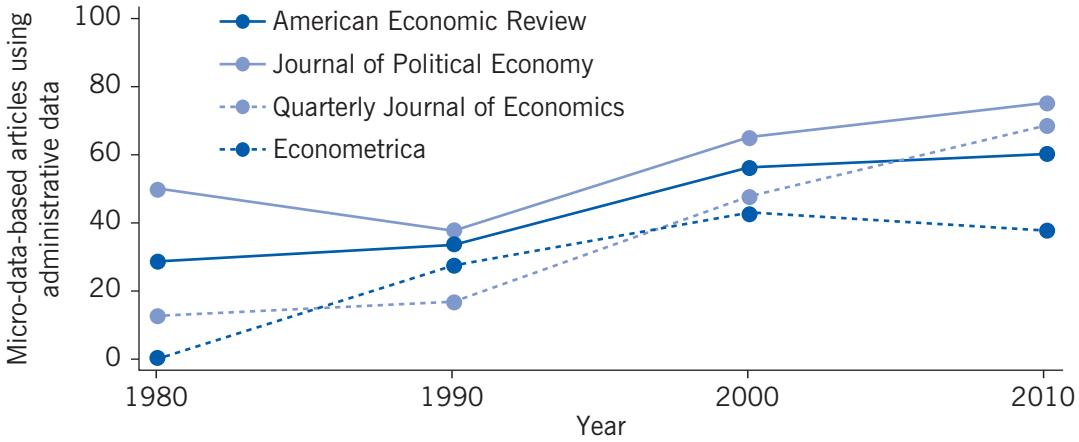
DISCUSSION OF PROS AND CONS

Administrative records

The advantages of administrative records as a data source for scientific research are straightforward [1]. Administrative records are consistently and accurately collected, resulting in highly reliable data covering a large number of observations, in some cases even 100% of the population (for example, in Scandinavian countries). They are regularly updated, so that the same data elements are typically available for long periods of time. In addition, the provision of administrative data for scientific research reflects a cost-effective way of providing highly reliable and representative data, as these data are already being collected for administrative purposes. Finally, many countries have begun to open their administrative databases for scientific research. As a result, the use of administrative data for empirical research in economics has increased dramatically since 1980. Figure 1 shows a clear rise in the number of micro-data-based studies published in top economic journals between 1980 and 2010 that used administrative records data.

However, there are also some limitations associated with administrative records data, which can reduce their usefulness for scientific research [1]. Access is often strictly controlled because of data security concerns, and the range and variety of variables is limited since administrative data are collected for administrative purposes rather than for research purposes. Variables of growing importance in empirical studies in labor economics, for example, such as social networks, personality traits, cognitive skills, attitudes, and ethnic identity, are usually not of interest to the administrators who collect the data and thus the information is not included in these databases [2]. Moreover, the way certain variables are recorded in administrative records (again for administrative purposes) sometimes diverges from fundamental economic concepts. For instance, recording monthly working income alone, without the underlying hours worked, might be sufficient in administrative data collected for tax purposes (the main purpose for collecting the data). For economists,

Figure 1. The use of administrative data in economic research has been rising, 1980–2010



Source: Chetty, R. *Time Trends in the Use of Administrative Data for Empirical Research*. Cambridge, MA: NBER Summer Institute, July 2012.

however, not having data on both income and hours worked makes it impossible to calculate hourly working income, an important statistic in labor economics. As a result, the information included in administrative records can often be used only as a proxy in scientific research or may not be useful at all.

Research potential of linked data

Linking administrative records data with survey data helps to overcome the shortcomings of administrative data discussed above by enhancing the usually limited set of information recorded. Not only does this linkage increase the number of variables, but the information that is already routinely collected for administrative purposes can be complemented by additional survey information to capture fundamental economic concepts in a way that is more amenable to economic research.

Take, as an example, the evaluation of active labor market policies, such as training and job search and placement activities. One of the main advantages of the linked data is that administrative records provide detailed and precise information about the labor market program in which individual unemployed workers have participated. By comparison, self-reported survey information on program participation is usually associated with measurement error, as respondents are often uncertain of the exact name of the program in which they participated or may not even be aware that they participated in a program, as in the case where a wage subsidy is paid directly to the employer. Combining the two data sources creates a much richer database. The highly reliable information on program participation and labor market biographies that is extracted from the administrative records data is thus enriched by important additional information from the survey data. The availability of survey information (such as details on personality traits, attitudes, expectations, and job search behavior) in addition to rich administrative data enables more detailed analysis of selection into programs and the determinants of labor market outcomes, as well as estimation of the effects of the program [3]. This example illustrates the need for linked data sets in empirical research in labor economics and its great potential for expanding analytical possibilities and accuracy.

Not only can survey information complement administrative data, but the combined data set can have strong positive impacts on the survey. Data linkage can reduce the costs of survey data collection by requiring fewer questions during interviews, since sensitive information such as wages can be taken directly from administrative records (see [4] for a practical example). Supplementing survey data with administrative records data reduces the burden on respondents and should, as a result, lead to higher survey participation [5]. The data linkage can also provide new insights into methodological questions on the validation of survey information, such as the reliability of self-reporting and the incidence of non-response bias [6].

Although there are many clear advantages to linked administrative records and survey data, in practice, the process of linking the data can be complex, costly, and time-consuming. These problems are due mainly to technical challenges in linking the data and the need for lengthy negotiations with data providers and government institutions motivated in large part by privacy concerns [7]. These obstacles need to be resolved before the data sets can be linked and used for scientific research. The following sections discuss the main challenges and provide an overview of practical solutions that have been implemented to overcome them.

Technical challenges

Linking administrative records data with survey data requires that the two data sets contain overlapping information. The most straightforward situation occurs when both data sources contain a unique personal identifier, such as a social security number. In practice, this is most often the case when the survey population was drawn from the administrative records, with the intention of matching the data later on. In this ideal case, the data can be directly linked, usually with almost no error. In the literature, this procedure is known as "exact matching."

In the absence of unique identifiers, combinations of other available individual characteristics have to be used instead, such as name, gender, address, and date and place of birth, to identify identical subjects in both data sets. In the literature, these characteristics are called "imperfect identifiers" as they may not be unique, and they may vary over time (addresses and names change). Sometimes one or the other data set will contain typos or other logging errors, or inputs may be missing entirely. Matching two data sources based on imperfect identifiers is thus not straightforward; it requires a multi-step, iterative process [8]. The first step is generally to standardize the imperfect identifiers by cleaning up the data, for example, by replacing nicknames with full names, standardizing the use of diacritical marks and upper and lower case, and fixing typos. Next, a similarity measure has to be created to capture the similarity between names, addresses, and other characteristics. Finally, the multiple similarity measures have to be combined to determine whether the cases are matches or not.

In the past, researchers used to simply sum up the similarity measures or just match identical cases deterministically. Recently, more elaborate weighting methods have been developed, which assign different weights to different measures based on a statistical matching method. This is called "probabilistic record linkage," and statistical software packages are now available to perform such tasks (see [8] for an overview). The matching procedure can be very time-intensive and, depending on the size and quality of the data set, can take up to several months to complete.

The way to avoid most of these technical problems is to plan ahead. When creating new surveys, it should be standard practice to incorporate a personal identifier that is also in the administrative records to which the survey can be linked. It is worth spending some time on this issue before data collection, as it will save much more time in the end and yield better quality data when the survey data are eventually linked to administrative records.

Legal challenges and resulting data restrictions

The main challenge when linking administrative and survey data is to deal with privacy concerns and the data restrictions resulting from them. Privacy concerns are justified and necessary, as information in administrative records are collected as part of administrative processes that are usually conducted without the explicit agreement of the individuals involved. That means that the individuals whose information is collected never consented to the use of their administrative records for scientific research. The easiest way to deal with such privacy concerns is to inform survey respondents about the intention to merge survey and administrative data, along with any associated risks, and to ask for permission to use the collected survey data in such a way.

A problem with this approach, however, is that consent rates vary widely, with one study finding variations ranging from 19% to almost 97% for different data sets [9]. Such differences introduce consent bias into any empirical analysis that uses these data because individuals who give permission to link survey and administrative data sets likely differ systematically from individuals who deny consent. That study also offers evidence on potential mechanisms driving the consent behavior of respondents. Figure 2 summarizes the study hypotheses and findings. Some of the main drivers of the consent decision are fear that data are too personal, a general resistance to the survey, and individual inattentiveness.

Other studies have documented that a willingness to consent to the use of linked survey and administrative data is correlated with observable individual characteristics, with results that are not entirely consistent across studies (see, for example, [10]). For instance, some studies find age, gender, and income to be positively correlated with consent, while other studies find exactly the opposite or find no effect at all (see [9] for further references).

In sum, the evidence clearly shows that asking individuals for permission to link their data is likely to induce consent bias in most cases, which has to be taken into account when drawing inferences from a linked data set. However, it also has been documented that consent-bias is usually smaller than other sources of bias in surveys, such as measurement error or non-response [11]. To determine how to increase consent rates and reduce potential consent bias, one study estimated the importance of factors that are under the control of survey designers [6]. It found no influence for the wording of the consent question but a significant impact for the placement of the question in the questionnaire (questions at the beginning of a survey achieve higher consent rates) and characteristics of the interviewer (interviewers who would give consent themselves are also more likely to win the consent of respondents).

Some countries have legislated legal exemptions to consent for data use for scientific purposes, which creates legal space for linking data. Thus, even where an individual has not explicitly consented to the linking of the data, such legal exemptions allow linking to occur. An example would be when survey designers had not intended to link the survey

Figure 2. Mechanisms of consent behavior

Mechanism	*Description*	*Hypothesized outcome*	*Empirical evidence*
Uncertainty	Respondents give consent because they cannot recall the required information—they allow the requested information to be drawn from administrative records.	Consent	No
Confidentiality concerns	Respondents fear that administrative records contain sensitive personal data and might be misused.	Denial	Strong support for negative relationship
Resistance to the survey	Respondents have a general resistance to the survey (they distrust the institution behind the survey), have no interest in the survey, or otherwise are opposed to the survey.	Denial	Support for negative relationship
Relationship with the data-owning agency	Respondents have a relationship with the institution that provides the administrative records; for example, they receive benefits or services from the institution.	Ambiguous	No support for a direct relationship, but the relationship to a government institution has positive effects
Attentiveness	Respondents are impatient or inattentive and are not willing to read or understand the consent form.	Ambiguous	Support that inattentive respondents face a higher probability of giving consent
Interviewer effects	The characteristics or views of the interviewer may influence respondents' consent behavior.	Ambiguous	Little support

Source: Sakshaug, J. W., M. P. Couper, M. B. Ofstedal, and D. R. Weir. "Linking survey and administrative records: Mechanisms of consent." *Sociological Methods & Research* 41:4 (2012): 535–569 [9].

IZA World of Labor

data with administrative records at the time they constructed the survey and so did not request consent from respondents. Legal exemptions usually require that researchers demonstrate that the importance of the scientific inquiry for which linked data are being requested outweighs any related privacy concerns. If this legal condition can be met, researchers may receive permission to access and link the data for scientific study. In practice, however, this approach involves massive amounts of uncertainty since the ruling on whether the data can be used in this way usually depends on subjective assessments by lawyers or data security officers rather than on the strength of the researchers' arguments for scientific need.

Since individuals do not explicitly agree to the use of their administrative records for scientific research, it is clear that it becomes much more complicated for researches to gain access to such data if the associated survey lacks a consent question asking permission to link an individual's survey and administrative data. The approval process for linking data under such conditions of missing consent usually involves long negotiations with government institutions and data protection officers. The way to avoid time-consuming

negotiations and long delays is to ensure that surveys that are intended to be matched to administrative records include a question on consent to the use of linked data, which should be asked (preferably) at the beginning of the interview and by interviewers who are more likely to agree to such use themselves.

In addition, the high risk of identifying an individual's personal details in the linked data set usually means that a high degree of data anonymization is required, which severely restricts access to the data and reduces its research potential. Such strong privacy protections may be needed because survey respondents do not have full information about precisely which data are included in the administrative records. The same problem arises even if individuals give their consent for use of linked data. Therefore, the manager of the administrative data (typically a government institution) usually requires the exclusion or aggregation of information, to mask its identifiable personal properties, and may further restrict data access to protect individuals against data misuse.

There are trade-offs between preserving the research potential of the data and controlling data availability that need to be resolved through compromise. One common practice is to provide different versions of the linked data set to the scientific community. For instance, users could choose between downloading a publicly available scientific-use file that contains a limited set of variables or negotiating restricted access to a much richer data set containing more sensitive variables, by agreeing to data use in safe environments (on-site use) or through remote access. Remote access is the most restrictive type of data access. Researchers have no direct access to the original data but have to prepare their analysis routines based on artificial test data instead. The routines are then sent to the data provider, who replaces the test data with the real data and returns the anonymized outputs to the researcher.

Although there are good practical solutions for research based on remote access, from a scientific perspective it is much more desirable to have direct, physical access to the linked data, either through publicly available scientific-use files or through on-site use. Having direct access allows the researcher to study and understand the data structure in detail and thus to exploit the full potential of the data. For instance, while test data can be used to simulate marginal distributions of variables, joint distributions are hard to replicate in test data. Therefore, remote access can be considered a second-best alternative and is additionally very time consuming.

Examples of linkage projects of key data sets in labor economics

Figure 3 lays out some examples of data-linking projects that have been conducted in different countries using data sets that play an important role in labor economics research. Such data-linking initiatives range from projects based on fairly narrow samples, to those based on large representative surveys that are heavily used in labor economics, such as the US Current Population Survey, the German Socio-Economic Panel, and the Spanish Labor Force Survey. Figure 3 further reflects the rising interest in linking administrative records and survey data, especially in Europe, which is driven primarily by the improved availability of survey data and administrative records. Finally, while pure survey data are usually publicly available for scientific research, Figure 3 indicates that most linked data sets are either not publicly available at all or available only in safe environments or through remote access. Again, such restrictions arise from privacy concerns related to the use of administrative data.

Figure 3. Examples of research projects in labor economics using linked data sets

Country	*Year (start)*	*Project description*	*Access type*	*Source*
Germany	2004	Links data from the Survey of Health, Ageing, and Retirement in Europe with administrative data from the Research Data Center of the German Pension Insurance	Scientific-use file	Korbmacher and Czaplicki (2013)
Germany	2005	Links panel data from the Labor Market and Social Security survey on welfare recipients with administrative data on employees (employment and unemployment) from the Institute of Employment Research (IAB)	On-site or remote	Antoni and Bethmann (2014)
Germany	2012	Links panel data from Working and Learning in a Changing World, which contains population-representative information on schooling, training, family formation, and regional mobility, with administrative data on employees (employment and unemployment) from the IAB	On-site or remote	Antoni and Seth (2012)
Germany	2013	Aims to link a subsample of the German Socio-Economic Panel on immigrants with administrative data on employees (employment and unemployment) from the IAB	Not yet available; planned on-site or remote access	Brücker et al. (2014)
Germany	2013	Links survey data on employers with administrative data on employees (employment and unemployment) from the IAB	On-site or remote	Heining et al. (2013)
Germany	2015	Aims to link data from a representative survey on entries into unemployment (IZA Evaluation Data Set Survey) with administrative data on employees (employment and unemployment) from the IAB	Not yet available; planned: scientific-use file	Caliendo et al. (2011)
Spain	2006	Links data from the Spanish version of the European Labour Force Survey with administrative data from the social security and tax databases to obtain reliable information on wages	Scientific-use file	Martinez and Galindo (2011)
UK	2001	The Improving Survey Measurement of Income and Employment initiative, which is investigating measurement error in survey data, links survey data from a UK sub-sample of the European Household Community Panel with administrative data on employers' records and government benefits from the Department for Work and Pensions	Scientific-use file	Jäckle et al. (2005)
US	1973	Links survey data from the Current Population Survey with administrative records of the Social Security Administration (SSA) and administrative tax information from the Internal Revenue Service	Non-public	Kilss and Scheuren (1978)
US	1990	The New Worker–Establishment Characteristics Database, one of the first and largest matched employer-employee data sets in the US, links information from the decennial census with the Standard Statistical Establishment List	Non-public	Troske (1998), Troske et al. (2000)
US	1992	Links survey data from the Health and Retirement Study, a representative survey of households, with SSA data to study the economics, health, and demographic implications of retirement and aging	On-site or remote	Olson (1999)

Source: Detailed reference information can be found in the full reference list in the online version of the paper: http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data

IZA World of Labor

LIMITATIONS AND GAPS

The rising demand for linked data within the scientific community has intensified the need to find less time-consuming and restrictive procedures to access those data. The main challenge when linking administrative and survey data for scientific research is to find better and more reliable legal solutions that can simplify the process of matching and accessing linked data. Moreover, because linking these data sources is a relatively recent development, data providers may need to acquire more know-how with data provision before fully understanding the associated risks and becoming more experienced in determining which restrictions are necessary and which are not.

In addition to projects analyzing linked data on the national level, it would be interesting to link international data sources, such as German survey data and Spanish administrative data. In a world of increasing international migration, such an initiative could offer new research opportunities to labor economics, by enabling them to follow individuals across borders. However, besides the usual legal constraints that arise when linking national survey and administrative data, international linking presents another important challenge: how to track individuals in different data sources when they move between different countries.

SUMMARY AND POLICY ADVICE

Linking survey and administrative data offers the potential for many new research opportunities for scientific and policy-related projects. While the number of linking projects in labor economics has been growing, that number is still very small. Growth in the number of projects has been slowed by concerns for individual privacy since consent to share administrative data is rarely available unless obtained explicitly through surveys that request such permission from respondents. As a result, legal constraints on the use of administrative data limit access to linked data and reduce the number of variables available for analysis because of the need to anonymize the data. Issues of privacy and consent remain the main challenge when linking both data sources. To ease these bottlenecks, policymakers, who have a lot to gain from the findings of research using linked data, should facilitate data linkage projects for scientific research. Doing so would result in the more efficient use of existing records and could also spark new research projects that may contribute novel insights and allow for drawing more reliable policy conclusions.

Acknowledgments

The author thanks two anonymous referees and the IZA World of Labor editors for many helpful suggestions on earlier drafts. The author also thanks Jan Wergula for excellent research assistance.

Competing interests

The IZA World of Labor project is committed to the *IZA Guiding Principles of Research Integrity*. The author declares to have observed these principles.

REFERENCES

Further reading

Herzog, T. N., F. J. Scheuren, and W. E. Winkler. *Data Quality and Record Linkage Techniques*. New York: Springer, 2007.

Korbmacher, J. M., and M. Schröder. "Consent when linking survey data with administrative records." *Survey Research Methods* 7:2 (2013): 115–131.

Key references

[1] Arni, P., M. Caliendo, S. Künn, and K. F. Zimmermann. "The IZA evaluation data set survey: A scientific use file." *IZA Journal of European Labor Studies* 3:6 (2014): 1–20.

[2] Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel. "The economics and psychology of personality traits." *Journal of Human Resources* 43:4 (2008): 972–1059.

[3] Caliendo, M., R. Mahlstedt, and O. A. Mitnik. *Unobservable, but Unimportant? The Influence of Personality Traits (and Other Usually Unobserved Variables) for the Evaluation of Labor Market Policies*. IZA Discussion Paper No. 8337, 2014.

[4] Martinez, M., and J. O. Galindo. *Integrating Administrative Data into the LFS Data Collection*. Workshop on LFS Methodology Working Paper, 2011.

[5] Lane, J. "Linking administrative and survey data." In: Marsden, P. V., and J. D. Wright (eds). *Handbook of Survey Research*. Bingley, UK: Emerald, 2010; pp. 659–680.

[6] Sakshaug, J. W., V. Tutz, and F. Kreuter. "Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data." *Survey Research Methods* 7:2 (2013): 133–144.

[7] Calderwood, L., and C. Lessof. "Enhancing longitudinal surveys by linking to administrative data." In: Lynn, P. (ed.). *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley & Sons, 2009; pp. 55–72.

[8] Schnell, R. *Linking Surveys and Administrative Data*. German Record Linkage Center Working Paper No. 2013-03, 2013.

[9] Sakshaug, J. W., M. P. Couper, M. B. Ofstedal, and D. R. Weir. "Linking survey and administrative records: Mechanisms of consent." *Sociological Methods & Research* 41:4 (2012): 535–569.

[10] Kho, M. E., M. Duffett, D. J. Willison, D. J. Cook, and M. C. Brouwers. "Written informed consent and selection bias in observational studies using medical records: A systematic review." *British Medical Journal* 338:b866 (2009).

[11] Sakshaug, J. W., and F. Kreuter. "Assessing the magnitude of non-consent biases in linked survey and administrative data." *Survey Research Methods* 6:2 (2012): 113–122.

Online extras

The **full reference list** for this article is available from:
http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data

View the **evidence map** for this article:
http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data/map