25

Errors in Linking Survey and Administrative Data

Joseph W. Sakshaug[1,2] *and Manfred Antoni*[3]

[1] *Cathie Marsh Institute for Social Research, University of Manchester, Manchester, UK*
[2] *Department of Statistical Methods, Institute for Employment Research (IAB), Nuremberg, Germany*
[3] *Research Data Centre (FDZ), Institute for Employment Research (IAB), Nuremberg, Germany*

25.1 Introduction

Linking surveys with administrative records is regarded as a useful and cost-effective means of supplementing survey data collection with existing administrative sources. Evidence of the widespread practice of linking surveys with administrative data is reflected in Table 25.1, which shows a selection of large-scale surveys and the types of administrative data they link to. The types of administrative data linked to these and other surveys vary considerably, but most often include social security records, tax and benefit records, healthcare enrollment and billing records, and education records.

The basic idea behind this type of linkage is to bring together survey information and administrative information that belong to the same unit; in this context, a unit may represent, for example, an individual, family, household, or establishment. There are two general record linkage techniques that are commonly used for this purpose. Both vary in their level of complexity. The most straightforward approach is to match record pairs based on one or more unique identifiers (e.g., social security numbers (SSNs), establishment numbers, personal or tax identification numbers) that are common to both record sources. This approach is often referred to as exact (or deterministic) linkage. The second, and more complex approach, referred to as probabilistic linkage, is typically used when unique identifiers are not available, or are deemed to be unreliable. In probabilistic linkage, nonunique identifiers (e.g., first and last name, address, date of birth) are used to calculate likelihoods that a given pair of records belongs to the same unit. This information is then used to decide whether a record pair constitutes a likely match. Both techniques are regularly performed in survey applications involving linkage to administrative records.

There are several strengths of administrative records that make them attractive for linking to surveys. An important strength is that they are relatively inexpensive to obtain and use. The fact

Total Survey Error in Practice, First Edition. Edited by Paul P. Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West.

Table 25.1 A selection of large-scale surveys that link survey records to administrative records.

Name of survey	Country	Type of administrative data
45 and Up Study	Australia	Admitted patient; emergency department; cancer registry, mortality; medicare benefits schedule; pharmaceutical benefits scheme data
Australian Longitudinal Study on Women's Health (ALSWH)	Australia	Admitted patients data; cancer registry; perinatal data
Canadian Community Health Survey (CCHS)	Canada	Data on deaths; hospital records; medical services; mental health; work safety
English Longitudinal Study of Ageing (ELSA)	England	Hospital records; tax and benefit records
The Survey of Health, Ageing and Retirement in Europe (SHARE)	Germany	Social security data
Labour Market and Social Security (PASS)	Germany	Social security data
Health and Retirement Study (HRS)	United States	Social security data; medicare claims data
Panel Study of Income Dynamics (PSID)	United States	Social security data; medicare claims data
Scottish Health Survey (SHeS)	Scotland	Mortality records; hospital discharge; cancer records; birth characteristics
Understanding Society—The U.K. Household Longitudinal Study	United Kingdom	Tax and benefit records; education records; health records

that these data already exist also saves potential data collection costs as this additional information need not be collected via primary data collection. Furthermore, respondent burden is potentially minimized because there is no need for respondents to report information that may already exist in the linked administrative record. Second, administrative records can provide detailed longitudinal information about complex employment and medical histories as well as earnings and expenditure information that can be difficult for respondents to self-report with high accuracy. Therefore, the risk of measurement error due to misreporting can be lessened by utilizing such records. Moreover, many administrative databases are updated regularly (sometimes continuously) and the data are usually collected systematically with quality control checks performed.

The aforementioned advantages of linking surveys with administrative sources make them particularly useful for different types of scientific research, including both methodological and substantive. From a methodological perspective, the linkage of surveys with administrative records allows plentiful opportunities, including the validation of answers given by survey respondents that overlap with measures collected in the administrative data; the evaluation (and adjustment) of nonresponse bias if administrative records are available for both respondents and nonrespondents; the evaluation of undercoverage within a census and over/undercoverage in sampling frames; the discovery of under-reporting or over-reporting discrepancies in administrative databases; the examination of reidentification risks of microdata files; more parsimonious design of survey questionnaires; and updating of sampling frames (among other uses).

Linking survey and administrative information also offers many substantive research possibilities, including detailed microanalysis. Given that administrative records are used to administer public services, these records contain many outcomes measured with a high level of accuracy that would be difficult to achieve using survey self-reports, such as lifetime earnings, unemployment spells, medical expenditures, receipt of social benefits, or welfare payments. These outcomes, combined with survey data, are desirable to answer important research questions, including, for example, what is the impact of major illness on financial hardship? How much money would the social security system save by increasing the minimum retirement age? Do healthcare reforms affect medical spending?

While there is little doubt that linked survey and administrative records offer many scientific research opportunities, most users of linked data are unaware of the linkage methods used to combine the different data sources and the quality of the final linked product. Linkage is typically performed by technicians whose primary job is not research and who may not know for which purposes the linked data will be used. The linkage procedure itself involves several steps, some of which may involve subjective judgments and tradeoffs in which no optimal solution exists. Consequently, errors may be introduced through the linkage, potentially impacting the inferences and conclusions drawn from analysis of the combined data.

The purpose of this chapter is to examine various error sources that can be introduced during the linkage process. The remainder of this chapter is organized as follows. Section 25.2 provides a conceptual framework for the possible errors that can be introduced by linking survey and administrative records. Section 25.3 describes the linkage consent process, summarizes the literature on linkage consent issues, and reviews factors that may influence linkage consent rates and biases. Section 25.4 provides an overview of linkage using unique identifiers and summarizes the different errors that can occur. Section 25.5 examines linkage procedures that rely on non-unique identifiers and reviews the potential errors that can arise. Section 25.6 discusses applications of record linkage and offers some practical guidance. Section 25.7 concludes with a brief summary of the main take-home points from this chapter.

25.2 Conceptual Framework of Linkage and Error Sources

In this section, we outline a conceptual framework for the process of linking survey and administrative data that is useful for identifying possible errors that can occur during linkage. Figure 25.1 presents a visual depiction of five general stages surveys typically undergo prior to and during the linkage process. In the first stage, a sampling frame is acquired or created which contains a list of units (e.g., persons, households, establishments) that reflect the target population of interest. In the second stage, a sample of units is selected from this list using a random selection procedure. In the third stage, the survey field period begins and the sample of units is divided into groups of respondents and nonrespondents. Various errors can occur at each stage of this process, including under/overcoverage, sampling error, and nonresponse bias. Each of these error sources has been studied extensively in prior work (Groves, 2004; Groves and Couper, 1998).

The remaining stages in this figure are relevant to the process of linking surveys with administrative records. The fourth stage depicts the common situation in which respondents are explicitly asked for permission to link their survey information to corresponding administrative records. This stage is often referred to as the informed consent process. Many surveys are required to ask respondents for informed consent to link their records on the basis of country or local jurisdiction laws, and/or to fulfill stipulations set forth by research ethics committees. The informed consent step is an ethical procedure that can serve several purposes. One purpose is to inform respondents about the type of administrative data that is being requested. For example, in the United States, a common type of administrative information requested in surveys is

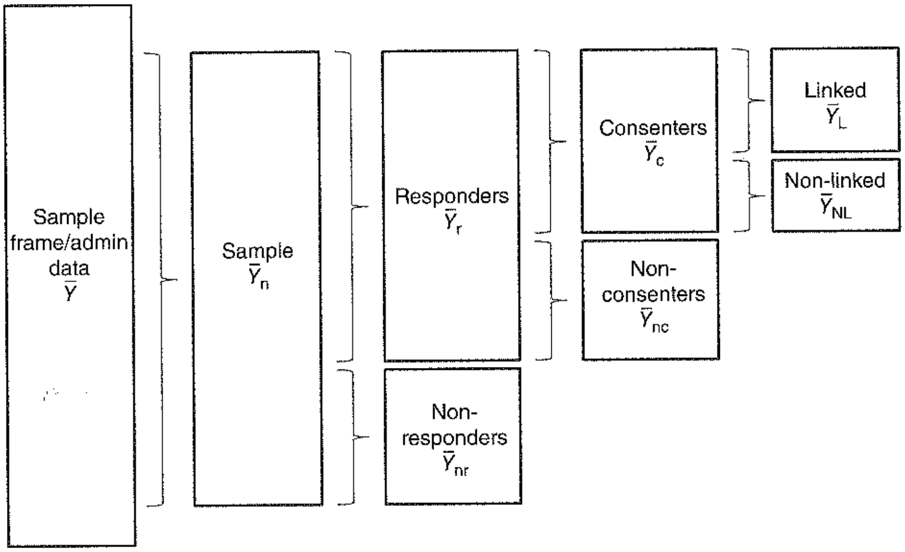
Sample frame/admin data $\bar{Y}$
Sample $\bar{Y}_n$
Responders $\bar{Y}_r$
Non-responders $\bar{Y}_{nr}$
Consenters $\bar{Y}_c$
Non-consenters $\bar{Y}_{nc}$
Linked $\bar{Y}_L$
Non-linked $\bar{Y}_{NL}$

Figure 25.1 Conceptual framework of the record linkage process.

earnings and social benefit histories contained in records housed at the Social Security Administration. A second purpose of the informed consent step is to describe the uses of the linked data and any potential benefits that may result. Another important purpose of the informed consent step is to alert respondents to the potential data confidentiality risks that may result from linking their data. This also presents an opportunity for the survey organization to describe the safeguarding procedures that will be put in place to minimize these risks. The actual implementation of the informed consent step and the degree to which each of the aforementioned purposes are fulfilled varies from study to study. Once the informed consent procedure is administered, respondents are then asked whether or not they agree to the linkage. If consent is provided then additional efforts to facilitate the linkage, for example, obtaining a unique identifier from the respondent (e.g., SSN) may be carried out.

For various reasons, not all respondents agree to link their survey information with administrative records. The error implications of nonconsent are twofold. First, the cases which do not consent to linkage are essentially dropped from the analysis of the linked data. Therefore, the analytic sample size is reduced and standard errors of estimates are inflated. The second possible implication is that systematic error in the linked-data estimates can arise if respondents who consent to the linkage are different from those who do not based on the linked variables of interest. Compared to the survey participation literature, the linkage consent literature is sparse and relatively few studies have examined linkage consent error, explored the mechanisms of its existence, and proposed methods of evaluating and adjusting for its occurrence.

The last stage of the conceptual linkage framework in Figure 25.1 describes the actual linkage process which takes place only for the subset of respondents who consent to record linkage. The actual linkage process typically unfolds using an exact or probabilistic linkage technique, or a combination of both techniques, for example, exact/deterministic record linkage is performed first and probabilistic linkage is subsequently performed for the remaining nonlinked cases. The final outcome of the linkage process is either a success or failure in terms of linking a particular survey unit with a corresponding administrative record. It is rare that all survey units eligible for linkage are successfully linked to their corresponding administrative record. Nonlinks can occur, for example, when an administrative record does not exist or cannot be found for a particular survey unit. Under the exact linkage procedure, a unique identifier obtained from the survey respondent may be incomplete or recorded with error and thus may not correspond to the same unit located in the administrative database. Alternatively, in the case of probabilistic linkage, the likelihood of a match for any given pair of records (based on the set of nonunique identifiers) may not meet or exceed the predefined threshold needed to classify the pair as a match. Just as linkage nonconsent reduces the analytic sample size of the linked database and may introduce bias, nonlinked cases (conditional on consent) also reduce the analytic sample size and may introduce bias if the nonlinked cases are systematically different from the linked cases.

Errors can also occur among the linked cases. For example, if the wrong unique identifier is provided by the respondent then it is possible that records belonging to two different units will be exactly linked (i.e., false positive match). Mistakenly linking records belonging to separate units is less common in exact linkage compared to probabilistic linkage. In probabilistic linkage, links are determined on the basis of nonunique identifiers and statistical likelihoods which make the linkage outcome more subjective compared to the simple deterministic rules of exact linkage, and more susceptible to false positives and false negatives. Furthermore, the likelihood calculations are based on several inputs provided by the linkage technician, such as the linkage weight for each identifier as well as thresholds which distinguish "likely matches" from "likely nonmatches." Therefore, different inputs provided by different linkage technicians can result in different false positive/negative rates.

The following sections provide a more in-depth examination of the aforementioned sources of linkage error: linkage nonconsent and erroneous linkage with unique and nonunique identifiers.

25.3 Errors Due to Linkage Consent

While linking administrative records to surveys is attractive from a scientific perspective, there are ethical and legal considerations that must be taken into account. One consideration is that augmenting administrative information about individuals could increase the risk of disclosure and the possible reidentification of survey participants. Such disclosures could potentially bring harm to individuals if the linked data are used for malevolent purposes. In such cases, not only could harm be brought to the individual but also legal action could be taken against the research institution responsible for collecting and safeguarding the data. A further consideration comes from a data privacy perspective. Individuals may not know exactly what information is contained in their record and may therefore prefer that this information remain confidential, or, even if this information is known to the individual, they may still have concerns about sharing such information with a third party.

Given the potential concerns over data privacy and misuse of the linked data, survey organizations are usually obligated to obtain informed consent from respondents before administrative information can be linked and released for research purposes. The process of obtaining linkage consent from respondents varies from study to study, but generally there are two key steps that occur. In the first step, respondents are presented with information about the linkage, such as the purpose of the linkage and the scientific objectives. Possible benefits (individual, societal, or otherwise) of the research may also be presented as well as any potential risks that may arise due to the linkage, and assurances of data confidentiality or a description of specific safeguards put in place to protect the data from unauthorized use. Often a form containing detailed information about these points is provided to respondents before or during the actual linkage request. The purpose of providing this information is to ensure respondents have a clear understanding of what is being requested. In the second step, respondents decide whether they agree or object to the linkage. The decision is usually documented with a signature, but in some surveys (e.g., telephone), verbal consent is sufficient.

Not all respondents agree to provide linkage consent and consent rates tend to vary widely across the social, health, and economic discipline areas. In a meta-analysis of health record linkage studies, da Silva et al. (2012) observed a range of consent rates between 39 and 97%. In another review of general linkage studies, Sakshaug and Kreuter (2012) reported a slightly wider range between 24 and 89%. Moreover, there are some indications that linkage consent rates are declining over time (Fulton, 2012). For example, between years 1993 and 2005, the linkage consent rate for the U.S. National Health Interview Survey declined from 85 to 50%. From 1996 to 2004, the consent rate from the U.S. Survey of Income and Program Participation (SIPP) fell from 88 to 65%. The U.S. Current Population Survey also experienced a decline from 90 to 76% between 1994 and 2003 (Bates, 2005).

25.3.1 Evidence of Linkage Consent Bias

Failure to achieve 100% consent rates can reduce the quality of the linked data in at least two ways. First, low consent rates can reduce the sample size of the linked records and, in turn, increase the variance of the estimates, and, second, the estimates may be biased if respondents who consent to the linkage are systematically different from those who do not based on the study variables of interest. The latter point is a primary concern among researchers and has led to investigations into the extent of consent bias. Given that linkage consent is asked of survey respondents, nearly all of the collected survey variables can be used to identify correlates of consent. Some of the most common correlates of consent include, age, gender, education, income, foreign citizenship, health outcomes, and benefit receipt (Knies and Burton, 2014; Mostafa, 2016; Sakshaug et al., 2012; Sala et al., 2012). Thus, there are indications that these variables are susceptible to consent bias.

However, it is important to point out that many of these correlates vary in their direction and magnitude from study to study. For instance, age is a frequent correlate that shows both positive and negative associations with consent (Banks et al., 2005; Dahlhamer and Cox, 2007; Jenkins et al., 2006; Young et al., 2001). The number of health conditions also tends to have a contradictory effect between studies (Dahlhamer and Cox, 2007; Haider and Solon, 2000; Young et al., 2001). These inconsistencies may be partially explained by the diversity of study populations and types of administrative data linkages that were performed. The key conclusion here is that while the potential for consent bias exists, the direction and magnitude of the bias are study-specific. For studies that link to multiple administrative databases, such bias can be linkage-specific as respondents have been shown to consent at different rates for different administrative linkages requested within the same survey (Al Baghal et al., 2014).

The aforementioned studies examine the potential for consent bias in survey variables. However, what is also of interest to users of linked survey and administrative data is whether or not consent biases exist in the administrative variables. Identifying consent biases in administrative variables is a more challenging task because administrative records for respondents who do not consent to linkage are not normally made available by administrative data custodians. However, exceptions have been made for purposes of studying linkage consent bias. For example, Sakshaug and Kreuter (2012) used the "Labour Market and Social Security (German abbreviation: PASS)," a longitudinal study in Germany that investigates the social processes and unintended consequences of labor market reforms, to investigate the extent of consent bias in administrative variables collected from a federal employment database. The PASS survey includes a benefit sample of persons who are currently receiving an income assistance benefit. This sample is drawn from administrative records belonging to the German Federal Employment Agency. Using this register-based sample, the authors received approval to link the consent indicator variable from the survey to the administrative database to facilitate the comparison of administrative records belonging to the consenting and nonconsenting respondents.

Across six administrative variables, Sakshaug and Kreuter (2012) found statistically significant consent biases for two variables: age and foreign citizenship. Both were negatively related to consent, but the size of the biases was relatively small (<1 percentage point) compared to nonresponse and measurement error bias. Similar findings were observed by Sakshaug and Huber (2016) who implemented the same consent indicator procedure to estimate linkage consent bias in the German survey "Further Training as a Part of Lifelong Learning (WeLL)." Their findings revealed similarly small consent biases compared to nonresponse biases for cross-sectional and longitudinal administrative estimates.

25.3.2 Optimizing Linkage Consent Rates

Evaluations of linkage consent rates and bias have coincided with efforts to identify design features of the consent request that maximize consent rates in surveys. The majority of these efforts can be grouped into three categories: the placement of the linkage consent request in the questionnaire, the wording (or framing) of the request, and whether active ("opt-in") or passive ("opt-out") consent procedures are used. We review these design features in turn.

25.3.2.1 Placement of the Linkage Consent Request

Most surveys have traditionally asked for linkage consent at (or near) the end of the survey interview. The exact rationale for this strategy is unclear. Presumably, the conventional wisdom is that requesting access to relatively sensitive information contained in administrative records may provoke privacy concerns that threaten respondents' willingness to proceed with the interview. Anecdotally (though not empirically confirmed), it is thought that interviewer-respondent rapport reaches its peak at the end of the interview, which may make the consent request less off-putting to respondents. Regardless of the exact rationale, it is apparent that asking for consent to linkage has typically been regarded as a secondary interviewing task.

There has been very little experimentation into the effects of placement on linkage consent rates, but all available evidence suggests that asking for consent at the end of the interview is suboptimal. Sala et al. (2014) found that placing the consent question in the context of topic-related items (e.g., asking for consent to link employment records when other employment-related items are being collected) leads to higher consent rates than asking for consent at the end of the interview. In a simple beginning- versus end-placement experiment, Sakshaug et al. (2013) found that asking for consent at the beginning of the interview yielded a higher consent rate than asking at the end. The same placement result has since been replicated in other work (Kreuter et al., 2015).

25.3.2.2 Wording of the Linkage Consent Request

Aside from placement, another feature of the linkage consent request that is under the control of the survey designer is how the request is worded. There is some variation in how linkage consent requests are worded across surveys (for a selection of different wordings, see Sakshaug et al., 2013) and there has been interest in identifying whether specific wordings increase respondents' likelihood of consent more than others. One strategy that is sometimes adopted in linkage requests is to emphasize the positive benefits of linkage to respondents. The experimental evidence is mixed on the effectiveness of this strategy. Pascale (2011) found no effect on consent rates when each of the following linkage benefits was presented to respondents: improved data accuracy, reduced costs, and reduced respondent burden. Sakshaug et al. (2013) also found no effect of benefit wording when the time-saving element of linkage was emphasized ("In order to keep the interview as short as possible, we would like..."). Both of these studies were conducted over the telephone with an interviewer. In a self-administered web survey implementation, Sakshaug and Kreuter (2014) replicated the time-saving argument experiment and found it to have a positive effect on the consent rate. It is plausible that benefit wording has stronger effects in visually-centric modes than in audio-centric modes due to variation in the degree to which interviewers read and emphasize the particular benefits.

Given the mixed evidence of benefit wording, there has been some experimentation into whether emphasizing the negative consequences of not linking one's data is a more effective strategy. This strategy is motivated by prospect theory (Kahneman and Tversky, 1979, 1984), which basically demonstrated that people's decision-making is influenced by whether the available choices are framed in terms of gains or losses. Through a series of experiments, Kahneman and Tversky showed that people are risk averse when faced with sure gains and risk seeking when faced with sure losses. There is some evidence that "loss framing" leads to higher linkage consent rates in surveys. Kreuter et al. (2016) found in a telephone survey that framing the linkage consent request in terms of losses ("The information you have provided so far would be *much less valuable* to us if we can't link it to...") led to a 10 percentage point increase in the consent rate relative to the gain-framing group ("The information you have provided so far would be *a lot more valuable* to us if we could link it to..."). In a subsequent replication study, loss framing yielded a modest 3 percentage point increase in linkage consent in a telephone study and a 13 percentage point increase in a web study (Kreuter et al., 2015), again suggesting that wording effects tend to be stronger in self-administered modes.

25.3.2.3 Active Versus Passive Consent

Linkage consent procedures generally fall into one of two categories: active or passive. Active consent procedures require that respondents explicitly express their willingness to consent either in writing or verbally. In other words, respondents must take some action in order to "opt-in" to the linkage, either by providing a signature and, in some cases, a personal ID number (e.g., SSN), or by simply providing oral confirmation. Most surveys adopt this explicit linkage procedure. In contrast, passive consent procedures require respondents to take action only if they do not agree to the linkage. That is, by taking no action (i.e., not opting out), respondents implicitly consent to the linkage. There has been considerable controversy in the survey participation literature on the merits of both active and passive procedures, but what is generally clear to all parties is that active procedures tend to yield lower rates of consent than passive procedures. However, evidence from the linkage consent literature is very limited. Das and Couper (2014) observed a very high linkage consent rate (about 95%) when an opt-out procedure was adopted. Bates (2005) found that respondents were less likely to consent to a hypothetical linkage request when explicitly requested to document their consent with an SSN, compared to respondents who were only asked if they objected to the linkage. Further evidence of the negative effects of asking for identifying information is provided by Sala et al. (2014) who reported in their study that about 4% of respondents agreed to linkage but were unwilling to provide a signature.

25.3.2.4 Obtaining Linkage Consent in Longitudinal Surveys

Linking administrative records to respondents in longitudinal surveys is quite common as it allows researchers to relate changes reported in the survey with changes in administrative circumstances over time. Linkage consent is typically requested repeatedly in longitudinal surveys, either from respondents who did not agree to linkage in the prior wave(s), or from all respondents in each wave including those who did agree to previous linkages. The former situation is relevant when consent is needed only once from respondents and the latter is relevant when consent agreements need to be renewed more frequently, e.g., at each wave. Strategies for maximizing consent rates in both situations have been investigated. For example, when linkage consent is needed only once over the course of a panel study, Sakshaug and Huber (2016) found that repeatedly asking for consent from respondents who did not provide consent in the previous wave not only increased the overall consent rate, but also reduced consent bias in subsequent waves. When consent is sought among all respondents in each wave, Sala et al. (2014) experimented with an independent/dependent interviewing approach. They found that reminding respondents that they had provided linkage consent in the previous wave (dependent interviewing approach) led to higher consent rates than not reminding them of their decision (independent approach). However, the dependent interviewing strategy tended to backfire for respondents who did not provide consent in the previous wave. That is, when reminded of their decision not to consent to linkage in the previous wave, respondents were less likely to give consent in the current wave compared to respondents who were not reminded of their previous refusal.

25.4 Erroneous Linkage with Unique Identifiers

Using unique identifiers to link records on a given observational unit from different data sources is the easiest and least error-prone method of data linkage. Common unique identifiers relating to people are SSNs, tax identification numbers, health insurance numbers, and personal identification numbers. For surveys of business entities, the linkage may rely on, for instance, establishment, company or health insurance company numbers. A common trait of most of these numbers is that they are issued by public bodies, though with varying degree of centralization across different countries. What makes these identifiers even more attractive for data linkage purposes is that their structure is strictly regulated. Aspects subject to such regulations are, for instance, how many characters these identifiers may have, how many and at which position letters are allowed, and what information specific digits or letters at certain positions should represent.

Despite the advantages such identifiers offer, they also show some inherent shortcomings. On the one hand, some countries (such as the Nordic countries) have comprehensive person identifiers that are issued very early in a person's life, are time-consistent, and are used by a broad range of jurisdictions, if not even by all of them. On the other hand, in many other countries, these numbers are only valid within the sphere of a single public body, e.g., the social security system in the United States. Thus, when administrative records from different sources have to be linked to survey data, the process of making sure that the identifier is collected correctly may have to be repeated, thereby increasing the risk of errors. Another inherent shortcoming of such identifiers is that their assignment may be decentralized, e.g., in countries with a strong federal system. A lack of coordination may result in numbers being issued more than once. For that and other administrative reasons, numbers may not be time-consistent. This results in another shortcoming of such numbers: survey respondents may remember and report their originally issued identification numbers correctly, but still provide invalid numbers, because they are unaware that their individual numbers have been changed within the administrative records.

Besides possible inherent shortcomings of such numbers, their collection and utilization in surveys are subject to several other risks related to remembering or reporting the correct identifier. Straightforward examples would be that respondents simply do not or only partially remember a certain identifier, or they only think they remember it correctly but in fact report the wrong number. Errors of that sort are made more likely by the fact that such numbers usually are hard to remember, e.g., because they are rather long and some of them are needed very rarely in a person's life. In societies that do or did heavily rely on the male breadwinner model, a specific problem with older women may arise: they may never have been issued a social security or tax identification number as they have never before earned money from dependent employment. However, instead of reporting that no relevant number exists for them, they might report the respective number of their husband, especially when he is already deceased and his number may have been relevant for pension claims of the widow.

Even if these numbers are remembered correctly, they may still be entered incorrectly into the database used for linkage. Errors may occur at every step of the process of data capture: the respondent may transpose characters when they have to write the identifier down themself, the interviewer may not understand the respondent correctly, or the interviewer may incorrectly write down the originally correct answer. If identifiers are not originally captured electronically but written down on a piece of paper, clerks may later have problems reading a stranger's handwriting or may simply make a mistake when copying the information into a database. Another type of recording error may occur when the actual respondent provides their unique identifier number instead of the target respondent.

Besides these aspects of measurement error, asking for unique identifiers may also increase item nonresponse. For some surveys, unique identifiers are directly available in the sampling information, e.g., when the sample is directly drawn from the same administrative records that will later provide additional research data. However, it is more common that the sampling information only contains some or all of the nonunique identifiers discussed in the next section. In this case, interviewers may only have to ask for linkage consent, as the identifiers are already available within the sampling information. If a unique identifier is to be used that is not already known, interviewers also have to collect information that may be seen as very sensitive by less trusting respondents. This may cause respondents to not report the relevant identifier or, because the sensitive nature of the administrative data becomes more salient, to not consent to data linkage at all.

To reduce the likelihood of such problems, several measures can be taken at different stages of the survey process. Advance letters may inform respondents of the planned data linkage. That way, respondents may be asked to prepare for the actual interview by looking up the relevant identifier beforehand. One has to bear in mind though that announcing the consent request in such a letter may have adverse effects on unit nonresponse. Respondents should be given the chance to subsequently report the relevant identifier by phone, email, or online form, if they were willing but unable to provide the information during the actual interview. Capturing the relevant information electronically is always preferable to having it written down on paper. First, this avoids possible errors by data typists later in the process. Second, it allows the people responsible for the survey instrument to include checks on whether the information entered into a form has the proper structure or whether numbers have been given correctly. Some identification numbers, for instance, include rules on how and where the gender or the birth data of the person has to be coded into the number. Proper interviewer training is also important in this respect. Interviewers should be familiar with rules governing the structure of the identifier in order to be able to detect inconsistencies. They should also be made aware that, in some instances, a partially given identifier may still be helpful, e.g., to detect false positives after alternatively linking the data sources by means of nonunique identifiers. Finally, interviewers should be aware of who the target respondent is to avoid collecting linkage identifiers on the wrong person.

25.5 Erroneous Linkage with Nonunique Identifiers

When unique identifiers are not available or their collection suffers from high unit nonresponse or measurement error, record linkage on nonunique identifiers is the next logical option. This section explores common linkage identifiers, identifies the challenges their collection or usage might pose, shows international best practice applications, and provides practical guidance to avoid errors in linkage using nonunique identifiers.

25.5.1 Common Nonunique Identifiers When Linking Data on People

The most common identifiers used when linking data on people are first and last name, address, birth date, and sex. Another common identifier is a respondent's birth name, as the last name may have changed without that change being registered in the administrative records. Depending on the type of databases to be linked, additional identifiers may be, for instance, geocodes of residential or working addresses, hospital admission dates, or International Classification of Diseases (ICD) codes. Error sources common to most of these identifiers are misspellings, inconsistent abbreviations (Dr. vs. Doctor, Street vs. St. vs. Str.), or omissions (day or month of the birthday, second or third given name, house number from address). Some identifiers are subject to specific errors, such as nicknames instead of given names, transposed digits in zipcodes, or switched elements of birth date or name. The reader is referred to Christen (2012, pp. 42–48) for a comprehensive overview of error sources with names as matching variables.

25.5.2 Common Nonunique Identifiers When Linking Data on Establishments

When linking an establishment or company survey to administrative data, completely different nonunique identifiers are relevant, the most common ones being name, legal form, and address. Additional identifiers, sometimes only used to check the linkage result for false positives, may be the economic sector, the number of employees, or the number of plants within a country. Identifiers such as company names or addresses potentially suffer from the same error sources as comparable identifiers on people. Some errors, though, are specific to identifiers of business entities. The economic sector of an enterprise, for instance, may easily be subject to misreporting, as the respondent could derive her answer from the main product that is produced at the plant, from the service or product that creates the most revenue for the superordinate company, or from the occupations the majority of workers are working in. Finally, the rationale of assigning the industry code by the respondent at the business does not necessarily have to be identical to that of the public body holding the administrative data. To make matters worse, both may be subject to change over time as the person reporting the information, the rules regulating the decision, or the actual focus of the business may change.

Additional problems may arise when linking data on different hierarchical levels such as companies and their subordinate establishments, i.e., when creating a one-to-many relationship. First, relevant identifiers on different levels may vary due to deviating rules or conventions of data capture. For instance, while the company level data may provide information on the legal form, that information may be missing in the establishment level data. Conversely, records in the establishment level data may contain added information on the location or purpose of a plant. Both cases can reduce the quality of the linkage result if the variation is not dealt with successfully during the data cleaning step. Second, one could compare the number of employees in the company with the sum of employees over all linked establishments to assess the quality of the linkage result. However, if the company survey, by error of the respondent or due to imprecise questionnaire design, collected information on the number of employees of the whole international company group, establishment data on the national level would lead to an erroneous comparison. This hints at a general problem when the goal of a linkage is a one-to-many relationship: as long as the number of subordinate units is unknown (e.g., from the survey or a third data source), the linkage practitioner can never be sure how many of the existing subunits have been found. This makes the detection of false negatives very resource intensive if not impossible.

25.6 Applications and Practical Guidance

This section shows best practice examples of linkage applications from different countries and offers practical guidance for future linkage projects to avoid known sources of error. For a comprehensive overview of record linkage case studies with a strong focus on North American data sources, the reader is referred to Herzog et al. (2007, Chapter 3). Additionally, a broad range of application areas are covered by Christen (2012, pp. 11–22).

25.6.1 Applications

Publications with detailed descriptions of methods, problems, and solutions of record linkage projects are rather rare. Reports on linked datasets often restrict themselves to a description of the resulting data, including the number of linked observations or a comparison of the original sample with the final dataset. More sophisticated reports also elaborate on the methods applied and problems encountered. Even their length and degree of detail vary depending on the complexity of the linkage tasks and the matching variables used.

Card et al. (2004) link data from the SIPP with administrative data on the Medicaid program in California to validate Medicaid coverage within the survey data. They are fortunate to be able to use the SSNs reported by SIPP respondents to link survey and administrative data. Still, they report that some respondents either do not have a valid SSN or do not consent to the linkage of the two data sources. They describe how SSNs are collected, how the attempt is made to convince originally nonconsenting respondents to consent in later panel waves, and how SSNs are assigned to respondents that did give consent to linkage but did not provide an SSN (Card et al., 2004, p. 414). By comparing the original sample composition with that of respondents without an SSN, Card and his coauthors are able to shed light on how the imperfect linkage may influence the inference based on the linked sample. They clearly state that errors in the linkage process, i.e., when a survey respondent cannot be linked successfully, lead to an underestimation of Medicaid coverage (Card et al., 2004, p. 416). They also show that part of the differences in coverage rates between survey and administrative data can be ascribed to missing or invalid SSNs within the administrative system (Card et al., 2004, p. 417). The authors formally show how measurement error in Medicaid coverage affects the consistency of statistical estimators when using such data (Card et al., 2004, p. 418).

Having to rely on nonunique linkage identifiers to link survey data to administrative records of the German Federal Employment Agency, Schild and Antoni (2014) elaborate on their methods and challenges with a high degree of detail. They discuss their original matching variables, their data cleaning strategy, and the subsequent comparison, classification, and evaluation steps. Although this application is very specific to the German context, the text has been used as a template for a linkage workflow by a number of subsequent projects. As with all linkage projects conducted by the German Record Linkage Center (GRLC; http://www.record-linkage.de), the linked data contain variables related to the linkage process and success. This allows users of the data to determine which cases were linked in which step and with which level of certainty. With these kinds of metadata, researchers are able to either run robustness checks in substantive analyses or do methodological analyses regarding the data quality.

While the above studies discuss linkages of data on people, a number of authors report on linkage applications that combine company and establishment level data. The following studies provide ample information on initial structures and quality of their matching variables, how they conducted data cleaning, and the actual linkage of record pairs. Most importantly, they show the specific challenges of linkages between entities on different hierarchical levels.

Biewen et al. (2012) describe the project "Combined firm data for Germany" (KombiFiD). This project led to a linkage of data from three different data sources, namely the Institute for Employment Research (IAB), the National Statistical Office, and the Deutsche Bundesbank. One of the most challenging aspects of the enterprise thus was the restrictive German data protection regulations regarding a linkage of data across different institutional boundaries and without a uniform legal basis. Another important aspect of KombiFiD was that most of the survey data were collected on the basis of the companies' legal obligation to respond. Although survey participation was obligatory, it was necessary to obtain linkage consent of the companies covered in the data. Biewen et al. (2012, pp. 364–366) show that this led to considerable nonconsent bias. The authors continue with a detailed description of the linkage process, which included both linkages with unique (establishment number) and nonunique identifiers (names and addresses of companies), depending on the datasets to be linked.

Schäffler (2014) links a company survey based on commercial business data to the administrative establishment data of the IAB. There is much to be learned from his detailed description of the challenges and methods. First, the text elaborates on how to evaluate the quality of the nonunique and error-prone identifiers *company name* and *legal form*. Second, Schäffler (2014) gives an overview of problems that arise from nonunique company names and how such problems may arise in a national administrative database. The examples all relate to the German language and institutional background.

Finally, there are a number of texts on business data linkage that do not involve linking any specific survey data, but provide useful practical advice. Wasi and Flaaen (2015) do not describe an actual linkage project. They rather demonstrate a set of methods and software tools that allow practitioners to link company or establishment level data. They cover all steps of the linkage process, i.e., preprocessing, linkage, and clerical review. While their advice is broadly applicable, their examples focus on North American naming conventions for addresses or company names. Another valuable resource regarding matching procedures for business data, although again without a specific survey application, is the text by Raffo and Lhuillery (2009). Their guidelines on common methods of identifying inventors or companies in patent databases like PATSTAT provide ample advice for linkage application in other fields. Finally, the text by Winkler (1995) is rich with practical advice on the linkage of business data. Again, the text covers the whole linkage process and is focused on North American names and addresses.

25.6.2 Practical Guidance

This section provides practical guidance drawn from the experience of the authors, from published best practice examples, and empirical findings in the literature.

25.6.2.1 Initial Data Quality

As shown above, nonunique identifiers can be subject to various sources of error. One important implication of this finding is that considerable resources should be invested in making sure that the initial quality of linkage identifiers is optimal. Several means to achieve that goal are of a technical nature: computer-assisted survey instruments should, for instance, include on-the-fly checking of the plausibility of entries (e.g., against switches of day and month of birth, implausible year of birth, correct structure of reported unique identifiers); the instrument should provide forms for data entry that do not truncate after too few characters. This aspect is also important when choosing the data storage format. The character encoding scheme should make sure to allow for any umlauts or other diacritical signs being commonly used in the country in which the survey is conducted. If academic or aristocratic titles or generational suffixes are to be retained during data capture, they should be stored in separate fields instead of the field of the first or last name. Other means of assuring a high initial quality of linkage identifiers rely on interviewer instructions and training. They should create awareness of the importance of data quality for the final research data. For instance, if the survey instrument does not provide separate fields for aforementioned academic and aristocratic titles, interviewers should be instructed to not include them into any of the other fields or to at least abbreviate them consistently. Interviewers filling in forms should be aware of possible problems affecting data quality caused by, for instance, nicknames or initials and should ask respondents to provide the full original names instead.

25.6.2.2 Preprocessing

None of these measures can assure that the quality of linkage identifiers is perfect. Especially with bad data quality, preprocessing is the most important factor for the success of linkage projects. In some situations, as much as "90% of the improvement in matching efficiency may be due to preprocessing" (Winkler, 2009, p. 370). In most cases, the expenditure of time for preprocessing often exceeds efforts of the record linkage itself (comparison, classification). Cleaning and parsing of the input files may make up for up to 75% of the effort within the linkage process (Gill, 2001, p. 31).

The seemingly trivial but important advice to be drawn from this is to plan for sufficient time and manpower in advance or at the beginning of a project. The above findings should convince practitioners that sufficient resources for preprocessing are crucial for achieving a good quality and level of standardization of identifiers and thus a good linkage result. However, Randall et al. (2013) caution against an overstandardization of identifiers. Their experiments show that reducing the variability of identifiers too strongly may lead to an increase in false positive matches, thus reducing the overall linkage quality.

This tradeoff shows another aspect of preprocessing as an important early step of the record linkage process. While practitioners may learn much from previous applications in terms of useful linkage methods or commonly applied string comparators, much of what determines the success of preprocessing can be very specific to the given project. The methods and programming strongly depend on the naming conventions and language of the country or the institution the data stem from. A sophisticated set of tools for cleaning or standardizing identifiers that is highly valuable and efficient in one context may suddenly become almost useless when dealing with a different data source. This, again, is meant to caution practitioners, especially when they deal with varying project contexts, that sufficient time and resources should be attributed to preprocessing. Comprehensive overviews on challenges and methods of data processing are provided in Christen (2012, pp. 39–67) and Herzog et al. (2007, pp. 107–114). Advice specific to preprocessing of identifiers in business data can be found in Winkler (1995, pp. 359–362).

25.7 Conclusions and Take-Home Points

This chapter sought to review possible kinds of error that can occur when linking surveys and administrative records with a focus on obtaining linkage consent from respondents and conducting linkages that involve unique and nonunique identifiers. We have reviewed the empirical literature and described situations in which errors can be introduced throughout the linkage process. Understanding these linkage errors is important as they can potentially impact the inferences that are drawn from linked datasets. As record linkage procedures are often performed by nonresearch staff, such errors are often unbeknownst to researchers who ultimately analyze the linked data.

We conclude this chapter by providing the following take-home points with an eye toward minimizing linkage errors.

- Linkage consent bias can occur when respondents who agree to the linkage are systematically different from those who do not based on the study variables used in the linked-data analysis. The existence and magnitude of consent biases vary from study to study. However, the magnitude of consent biases in administrative variables tends to be small compared to other survey errors (e.g., nonresponse, measurement).
- The following methods have been shown to be useful in maximizing linkage consent rates: (i) placing the linkage consent question at the beginning of the survey questionnaire or in the context of related survey items—both tend to perform better than requesting consent at the end of the interview; (ii) emphasizing the negative consequences of not providing linkage consent to respondents (loss framing) tends to yield higher consent rates than emphasizing the benefits of linkage. In general, wording effects tend to be stronger in self-administered modes of data collection than in interviewer-administered modes; (iii) passive consent ("opt-out") procedures tend to yield higher rates of linkage consent than active consent ("opt-in") procedures; and (iv) in longitudinal surveys, repeatedly asking for linkage consent among respondents who did not provide consent in previous waves has been shown to increase consent rates and decrease consent bias over time. Reminding respondents of their previous consent decision is optimal only when they provided consent in the previous wave.
- Both unique and nonunique identifiers have their own shortcomings and diverse sources for errors. Unique identifiers allow for a straightforward way of linking records from different sources, but they may be harder to come by. Nonunique identifiers may often be readily available from the sampling information or seemingly easy to collect, but they are much more prone to include errors than unique identifiers. If resources of the survey allow to do so, both kinds of identifiers should be collected to be able to use them iteratively.
- The quality of linkage identifiers is of utmost importance for linkage success. High quality can be achieved by both an optimized survey process (computer-assisted survey instrument with built-in plausibility checks, interviewer instructions and training raising awareness of data quality) and substantial data cleaning and standardization after the data collection process. The impact of both measures on linkage quality usually outweighs that of the choice of comparison method or functional parameters.

References

Al Baghal, T., Knies, G., and Burton, J. (2014). Linking administrative records to surveys: Differences in the correlates to consent decisions. Understanding society technical report no. 2014-09, Institute for Social and Economic Research, Essex. https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2014-09.pdf (accessed June 9, 2016).

Banks, J., Lessof, C., Taylor, R., Cox, K., and Philo, D. (2005). Linking survey and administrative data in the English longitudinal study of ageing. Presented at the Meeting on Linking Survey and Administrative Data and Statistical Disclosure Control, Royal Statistical Society, London, May 19.

Bates, N. (2005). Development and testing of informed consent questions to link survey data with administrative records. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3786–3793. https://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000234.pdf (accessed June 9, 2016).

Biewen, E., Gruhl, A., Gürke, C., Hethey-Maier, T., and Weiß, E. (2012). Combined firm data for Germany: Possibilities and consequences of merging firm data from different data producers. *Schmollers Jahrbuch. Journal of Applied Social Science Studies*, 132, 3, 361–377.

Card, D., Hildreth, A.K.G., and Shore-Sheppard, L.D. (2004). The measurement of Medicaid coverage in the SIPP. *Journal of Business and Economic Statistics*, 22, 4, 410–420.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Berlin: Springer.

da Silva, M.E.M., Coeli, C.M., Ventura, M., Palacios, M., Magnanini, M.M.F., Camargo, T.M.C.R., and Camargo, K.R. (2012). Informed consent for record linkage: A systematic review. *Journal of Medical Ethics*, 38, 10, 639–642.

Dahlhamer, J.M. and Cox, C.S. (2007). Respondent consent to link survey data with administrative records: Results from a split-ballot field test with the 2007 National Health Interview Survey. *Proceedings of the Federal Committee on Statistical Methodology Research Conference.* http://fcsm.sites.usa.gov/files/2014/05/2007FCSM_Dahlhamer-IV-B.pdf (accessed July 13, 2016).

Das, M. and Couper, M.P. (2014). Optimizing opt-out consent for record linkage. *Journal of Official Statistics*, 30, 3, 479–497.

Fulton, J.A. (2012). Respondent consent to use administrative data. Unpublished dissertation, University of Maryland. http://drum.lib.umd.edu/handle/1903/13601 (accessed June 9, 2016).

Gill, L.E. (2001). Methods for automatic record matching and linkage and their use in national statistics. Technical report, National Statistics Methodological Series no. 25, Office for National Statistics, London. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf (accessed June 9, 2016).

Groves, R.M. (2004). *Survey errors and survey costs.* Hoboken: John Wiley & Sons, Inc.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in household interview surveys.* New York: John Wiley & Sons, Inc.

Haider, S. and Solon, G. (2000). Non random selection in the HRS social security earnings sample. Santa Monica: RAND Corporation. http://www.rand.org/pubs/drafts/DRU2254 (accessed June 9, 2016).

Herzog, T.N., Scheuren, F.J., and Winkler, W.E. (2007). *Data quality and record linkage techniques.* New York: Springer.

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A., and Sala, E. (2006). Patterns of consent: Evidence from a general household survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 4, 701–722.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 2, 263–291.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39, 4, 341–350.

Knies, G. and Burton, J. (2014). Analysis of four studies in a comparative framework reveals: Health linkage consent rates on British cohort studies higher than on UK household panel surveys. *BMC Medical Research Methodology*, 14, 125.

Kreuter, F., Sakshaug, J.W., Schmucker, A., Singer, E., and Couper, M. (2015). Privacy, data linkage, and informed consent. *Presented at the 70th Annual Conference of the American Association for Public Opinion Research*, May 15, Hollywood, FL.

Kreuter, F., Sakshaug, J.W., and Tourangeau, R. (2016). The framing of the record linkage consent question. *International Journal of Public Opinion Research*, 28, 1, 142–152.

Mostafa, T. (2016). Variation within households in consent to link survey data to administrative records: Evidence from the UK Millennium Cohort Study. *International Journal of Social Research Methodology*, 19, 3, 355–375.

Pascale, J. (2011). Requesting consent to link survey data to administrative records: Results from a split-ballot experiment in the survey of health insurance and program participation (SHIPP). U.S. Census Bureau research report series (Survey Methodology), no. 2011-03. https://www.census.gov/srd/papers/pdf/ssm2011-03.pdf (accessed June 9, 2016).

Raffo, J. and Lhuillery, S. (2009). How to play the "Names Game": Patent retrieval comparing different heuristics. *Research Policy*, 38, 10, 1617–1627.

Randall, S.M., Ferrante, A.M., Boyd, J.H., and Semmens, J.B. (2013). The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making*, 13, 64.

Sakshaug, J.W. and Huber, M. (2016). An evaluation of panel nonresponse and linkage consent bias in a survey of employees in Germany. *Journal of Survey Statistics and Methodology*, 4, 1, 71–93.

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

Sakshaug, J.W. and Kreuter, F. (2014). The effect of benefit wording on consent to link survey and administrative records in a web survey. *Public Opinion Quarterly*, 78, 1, 166–176.

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

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

Sala, E., Burton, J., and Knies, G. (2012). Correlates of obtaining informed consent to data linkage respondent, interview, and interviewer characteristics. *Sociological Methods and Research*, 41, 3, 414–439.

Sala, E., Knies, G., and Burton, J. (2014). Propensity to consent to data linkage: Experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*, 17, 5, 455–473.

Schäffler, J. (2014). ReLOC linkage: A new method for linking firm-level data with the establishment-level data of the IAB. FDZ-Methodenreport (technical report) 05/2014, Institute for Employment Research, Nuremberg. http://doku.iab.de/fdz/reporte/2014/MR_05-14_EN.pdf (accessed June 9, 2016).

Schild, C.J. and Antoni, M. (2014). Linking survey data with administrative social security data—The project "Interactions between capabilities in work and private life." Working paper no. wp-grlc-2014-02, German Record Linkage Center, Nuremberg. http://record-linkage.de/-download=wp-grlc-2014-02.pdf (accessed June 9, 2016).

Wasi, N. and Flaaen, A. (2015). Record linkage using Stata: Preprocessing, linking, and reviewing utilities. *Stata Journal*, 15, 3, 672–697.

Winkler, W.E. (1995). Matching and record linkage. In B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds) *Business survey methods*, 355–384. New York: John Wiley & Sons, Inc.

Winkler, W.E. (2009). Record linkage. In D. Pfeffermann and C.R. Rao (eds) *Handbook of statistics 29A, sample surveys: Design, methods and applications*, 351–380. Amsterdam: Elsevier.

Young, A.F., Dobson, A.J., and Byles, J.E. (2001). Health services research using linked records: Who consents and what is the gain? *Australian and New Zealand Journal of Public Health*, 25, 5, 417–420.