Survey Research Methods (2017)
Vol. 11, No. 1, pp. 63-80
doi:10.18148/srm/2017.v11i1.6718

ISSN 1864-3361
http://www.surveymethods.org

The Quality and Selectivity of Linking Federal Administrative Records to Respondents and Nonrespondents in a General Population Survey in Germany

Joseph W. Sakshaug
School of Social Sciences
University of Manchester
Manchester, United Kingdom

Manfred Antoni
Institute for Employment Research
Nuremberg, Germany

Reinhard Sauckel
Independent Researcher
Munich, Germany

Various forms of auxiliary information are being sought to augment general population survey samples in order to evaluate and improve the representativeness and overall quality of survey data. However, auxiliary data options are limited in most general population surveys. Federal administrative databases provide a potentially rich source of auxiliary information, but linking them to general population samples is often restricted to surveys which draw their samples from population registers containing unique identifiers which can be directly linked to federal databases. In this article, we examine the quality and selectivity of linkages between a general population survey sample and a federal administrative database performed without a unique identifier. We employ a series of standard linkage procedures that rely instead on non-unique and error-prone identifiers obtained from the sampling frame to link a federal employment database to respondents and nonrespondents in a nationally-representative survey in Germany. The quality and selectivity of the established links are evaluated using sample disposition codes, and household- and person-level interview data in accordance with German data protection laws. We report a linkage rate of 60 percent for the entire sample under a strict linkage criterion, and 80 percent under a more relaxed criterion. We find that linkage rates vary across some sample disposition codes as well as household- and person-level characteristics that are likely specific to the particular administrative database used in this case study. We conclude with a general discussion of the practical implications of this work for survey organizations considering performing similar linkages and highlight some opportunities for further research.

Keywords: paradata, administrative data, auxiliary data, record linkage

1 Introduction

Survey organizations are actively looking for sources of auxiliary data to improve survey operations, increase cost efficiencies, and enhance survey estimation. Moreover, the need for effective auxiliary information coincides with a host of survey methodological activities, including the use of sample representativeness indicators (Schouten, Shlomo, & Skinner, 2011; Wagner, 2012), the evaluation of mode effects in mixed-mode surveys (de Leeuw, 2005; Vannieuwenhuyze & Loosveldt, 2013), and the implementation of responsive survey designs (Groves & Heeringa, 2006). However, for most general population surveys, auxiliary data options that permit these activities are limited (for an extensive review, see K. Olson, 2013). One of the most common sources of auxiliary information come from process-oriented paradata collected during the sample recruitment (Couper, 1998; Couper & Lyberg, 2005; Kreuter, 2013). These data include, for example, the date and time of each contact attempt, the data collection mode used for each attempt, and whether a message was left, incentive offered, etc. A second source of auxiliary data come from interviewer observations, where interviewers are instructed to make observations about the sampled neighborhood, housing unit, or members of the housing unit that are likely to be related to key survey variables (Diez Roux, 2001; Peytchev & K. Olson, 2007; West, 2013; West, Kreuter, & Trappmann, 2014).

Contact information: Joseph W. Sakshaug, School of Social Sciences, University of Manchester, Bridgeford Street-G12, Manchester M13 9PL, United Kingdom, email: joe.sakshaug@manchester.ac.uk.

Other auxiliary data sources that have been explored include aggregate and geographically-coded data, commercial data, and public records. Smith and Kim (2013) identify 20 aggregate- and geocoded-data sources in the United States that can be linked to housing units (and their surrounding areas). Commercial databases can be purchased to obtain demographic, occupation, and financial information for households and their occupants. However, the availability, completeness, and quality of their content can vary considerably (DiSogra, Dennis, & Fahimi, 2010; Pasek, Jang, Cobb, Dennis, & DiSogra, 2014; Raghunathan & Hoewyk, 2008; Sinibaldi, Trappmann, & Kreuter, 2014; West, Wagner, Hubbard, & Gu, 2015).

Online "reverse directories" and other address-searchable databases have also been used to collect a range of information about sampled housing units and their occupants. Efforts to merge these databases to population samples are ongoing. At the National Opinion Research Center, a large auxiliary data project, referred to as the "Multi-Level Integrated Database Approach (MIDA)", makes use of multiple address-searchable databases to retrieve auxiliary information on sampled addresses. In their evaluation of MIDA, Smith and Kim (2013) found that at least some information could be merged to 93.5 percent of all sampled addresses, though merge rates varied between specific data fields. The authors note several limitations of these databases, among them: 1) their content may not be harmonized across jurisdictional boundaries; 2) data access restrictions vary across jurisdictions; 3) they do not always contain the most up-to-date information (e.g. names of people no longer living at the address) and; 4) item missing data rates can be rather high.

Another source of auxiliary information – one that we consider in this article – comes from federal administrative databases. While these databases are subject to many of the same problems mentioned above (e.g. item missing data, access restrictions), they possess some desirable properties that make their use as survey auxiliary data quite attractive. For instance, because their primary purpose is for monitoring and administering services to the population, they are usually updated regularly to reflect changes in an individual's program membership and eligibility status, and generally provide reliable longitudinal information about most population members. A second property is that their content is usually harmonized across local jurisdictions; thus, measurement differences are minimized across local reporting authorities. A third property relates to their process-oriented nature which makes their acquisition costs relatively low compared to primary data collection. Lastly, an important property of these data is that they often contain substantive variables that are related to common survey topics (e.g. healthcare utilization, expenditures, employment history, earnings, benefit receipt), which makes them a potentially useful source of auxiliary information for surveys.

Most applications of linking federal administrative records to surveys are motivated from a substantive rather than a methodological perspective. Such records are primarily used by researchers to study complex policy-oriented research questions, which are difficult to answer using survey data alone. Studying policy-relevant topics, such as labor market participation, healthcare spending, or poverty, is facilitated by merging survey responses with relevant administrative information concerning lifetime earnings, welfare participation, (un)employment spell durations, or direct healthcare costs. Recognizing the increasing demand for administrative information in substantive applications (Chetty, 2012), many large-scale surveys supplement their primary data collection activities with linkages to federal databases (Antoni & Seth, 2012; Freedman, McGonagle, & Andreski, 2014; Knies & Burton, 2014; Korbmacher & Czaplicki, 2013; J. A. Olson, 1999). These linkages are usually restricted to survey respondents and only those who consent to linkage, which limits their usefulness for survey methodological and operations research. However, assuming the relevant legal and ethical issues can be addressed – a point which we come back to later – it is conceivable that surveys could extend their administrative data linkages to all sampled units, including nonrespondents, to supplement other, more common, sources of auxiliary information (e.g. paradata).

Merging federal administrative records to all sampled units (respondents and nonrespondents) is not new, and in some countries it is routine. Many European countries (e.g. Denmark, Finland, the Netherlands, Sweden) maintain population registers that are frequently used to draw samples for surveys (Blom & Carlsson, 1999; UNECE, 2007; A. Wallgren & B. Wallgren 2007). These registers contain variables recorded for virtually all population members, including name, address, date of birth, sex, marital status, country of birth, or taxable income. They also contain a unique identifier that can be directly linked to other administrative databases containing more detailed information on employment history, education, housing matters, and other attributes. However, in most countries, general population survey samples are not drawn from population registers that can be directly linked to administrative databases. Even for surveys that link administrative records to survey respondents, a unique identifier (e.g. social security number) is often requested to facilitate direct linkage. Extending the linkage to all sampled units, including nonrespondents, would therefore necessitate the use of non-unique identifiers and indirect linkage procedures.

The notion of linking federal administrative records to general population samples is appealing from a survey methodological perspective, but before this approach can be deemed useful it is important to answer several linkage- and quality-related questions. For instance, is it feasible to link entire survey samples to administrative databases without a

unique identifier? Are linkage rates similar for respondents and nonrespondents? Are linked and non-linked cases similar with respect to key attributes, or do they systematically differ in ways that might compromise the generalizability of the linked data? Such questions have recently been explored by the U.S. Census Bureau (Bee, Gathright, & Meyer, 2015). In their case study, responding and nonresponding households from the 2011 Current Population Survey (CPS)'s Annual Social and Economic Supplement (ASEC) were indirectly linked to 2010 federal tax records from the Internal Revenue Service. About 79 percent of CPS responding and 76 percent of nonresponding household addresses could be linked to at least one tax record. The selectivity of the linkage based on respondent ASEC data revealed that lower-income households (indicated by adjusted gross income) were less likely to be linked to tax records, likely reflecting households that did not file a tax return and were thus not included in the record base. Other selectivities included households where the householder was older, less-educated, Black, Hispanic, never married, and non-U.S. citizens; such households were linked at lower rates compared to those with complement householder characteristics.

We extend this line of investigation in a European context by examining the quality and selectivity associated with linking a nationally-representative survey sample to a federal employment database in Germany. Unlike the case study by (Bee et al., 2015), our case study focuses on linkage at the individual level where names, addresses, sex, and a dichotomized indicator of birth cohort are available as part of the sampling frame for both respondents and nonrespondents. We acknowledge the availability of individual-level information for both respondents and nonrespondents is uncommon in many surveys. On the other hand, our case study is emblematic of many survey settings in which the target administrative database was not used as a sampling frame and not directly linkable to the survey sample via a unique identifier. We apply a series of standard, indirect linkage procedures and review the quality of the resulting linkage by examining linkage rates across sample disposition groups, and household- and person-level subgroups. Specifically, we conduct this study with an eye towards addressing the following research questions:

1. What proportion of the general population sample can be linked to administrative records using indirect linkage procedures?

2. How are the established links distributed across sample disposition codes and nonresponse types (noncontacts, refusals)?

3. To what extent do linkage rates vary across specific household- and person-level characteristics (available for the interviewed cases)?

2 Data sources and legal considerations

2.1 Survey data

The survey that serves as the basis for this case study is the German Labour Market and Social Security (PASS) study.[1] The PASS is an annual, longitudinal survey of households in Germany. The PASS was initiated in 2006 by the Institute for Employment Research (IAB) of the Federal Employment Agency (BA) in response to the country's reorganization of the welfare and unemployment benefits system – the so-called Hartz-reforms (Möller & Walwei, 2009). One of the major developments of these reforms was the introduction of a new benefit scheme called Unemployment Benefit II (UB II). UB II is a means-tested benefit that provides minimal assistance for individuals aged 15 to 64 who are able to work, but whose household has insufficient income. The PASS was launched to evaluate the consequences of this reform by collecting extensive information about the economic and social situations, behavior, and attitudes of benefit recipients, but it is also used to study poverty and deprivation more generally.

The PASS study is composed of two samples: a UB II benefit recipient sample and a general population sample. The UB II sample is drawn directly from administrative databases of benefit recipients at the IAB, whereas the general population sample is drawn from population lists compiled by municipality registration offices – resident registration is compulsory in Germany. Both population lists are stratified and each sample is drawn within selected primary sampling units. The UB II sample is refreshed with new entries to this population each year, whereas the general population sample is replenished less frequently. While both samples are composed of individuals, PASS is a household survey. A household interview is conducted with the person who is most knowledgeable about the household situation (who may differ from the sampled individual), and personal interviews are conducted with all household members starting from the age of 15 years. Everyone who moves or is born into a PASS household is included in the sample and followed after moving out. Data collection is conducted using a sequential-mixed mode design of computer-assisted telephone interviewing and computer assisted-personal interviewing. Further details about the PASS study design can be found in (Trappmann, Beste, Bethmann, & Müller, 2013).

For our purposes, we only link the general population sample of refreshment cases drawn for wave 5 to federal administrative records.[2] The refreshment sample consists of

[1] See http://fdz.iab.de/en/FDZ_Individual_Data/PASS.aspx for more details.

[2] The household interview response rate for the PASS wave 5 general population refreshment sample in accordance with AAPOR standards (AAPOR, 2016; Response Rate 1) is 24.5 percent. We use this particular PASS sample despite the fact that there have been

6,237 persons (and their corresponding households). Linkage is performed only on the sampled persons and not for other household members who were eligible to take part in the survey. The sampling information provided by the municipality offices does not include a unique identifier that can be used to directly link the sample to administrative records. We instead have to rely on the following non-unique and error-prone identifiers: *first name*, *last name*, *zip code*, *city*, *street name*, *house number*, *sex*, and dichotomized *birth cohort* (born before 1945 and after).[3] These are the only data that are part of the sampling frame and also available in the administrative data source, which we describe below. While the PASS study was conducted in 2011, we use administrative linkage identifiers from the years 2009 – 2012. The broader time range for the administrative data was intended to increase the linkage success.

2.2 Administrative data

We link the PASS wave 5 refreshment sample (and associated paradata) to administrative employment data of the IAB. Being part of the BA, the IAB has access to data originally collected by the BA for administrative purposes. The first major source of this administrative database is mandatory social security notifications of employers regarding their employees. Facts like start and end dates, wage sums, and the occupational status of job episodes are collected with high reliability and precision as they are relevant for unemployment and pension insurance entitlements (Antoni, Ganzer, & vom Berge, 2016; Jacobebbinghaus & Seth, 2007). The second source of these administrative data is longitudinal information on registered unemployment, job search, or participation of individuals in active labor market programs. These data are collected in a standardized manner in local employment agencies and are centrally stored and processed by the BA. The BA creates a consistent person identifier across these data sources, thereby allowing the IAB to derive longitudinal data that cover almost every aspect of a person's employment history. With all of these sources combined, the administrative data of the IAB cover most of the German working population. Excluded from the database are persons who are exempt from making social security contributions, including civil servants (e.g. teachers, professors, police officers), and the self-employed or homemakers.[4] We expect that the linkage will be less successful for these groups compared to others.[5]

An important distinction has to be made at this point between the personal identifying information (e.g. names, addresses) used for the record linkage and the research data (e.g. employment spells, income) used for analysis. Both stem from the same administrative notification processes and both have a longitudinal structure by nature. Due to strict privacy regulations, both elements of these administrative data are stored separately and are never used simultaneously in any of the steps of this project.

2.3 Legal considerations

What makes this project unique in the German context is that we are able to link sampled persons (and households) regardless of whether they consented to linkage or even participated in the survey. This unconsented linkage is only possible within the IAB because the IAB owns the sampling frame data from the survey (including names and addresses) and has access to necessary data from the BA to perform the linkage. While actual survey responses can only be linked to administrative data with respondents' consent, the survey paradata (and sampling frame details) only provide facts about the survey process and are not provided by individuals directly; thus, an argument can be made that these sample details can be linked to administrative data without violating one's data privacy rights.

Justification for this argument was sought through the IAB legal department. A statement describing the project and its research goals was submitted by the investigators and meetings with the IAB legal team took place to clarify the project details. The project statement included assurances on several conditions: 1) no survey information (beyond paradata) would be linked to the administrative database unless informed linkage consent was already obtained from the PASS respondents; 2) all personal identifying information (e.g. names, addresses) would be purged from the linked database after the linkage was complete; and 3) the linked data would only be used internally for research purposes and would not be disseminated outside of the IAB. The IAB legal department approved the project under these conditions.[6]

---

several additional panel waves afterwards. The reason is that wave 5 was the second and last wave so far with a large refreshment sample of the general population. Every wave since has included only previous panel members or small refreshment samples from the UB II population, directly drawn from the administrative data of the IAB.

[3]Instead of yearly birth cohorts, the sampling information only contained these two birth cohort groups. This is why we could not make use of the much more detailed birthdate information given in the administrative data.

[4]According to figures from the Federal Statistical Office for 2011 based on Microcensus data, these exclusions comprise about 12.5 percent of the total population of Germany aged 15 to 65.

[5]The linkage probability of people in these non-registered job types is still larger than zero as they may have held jobs subject to social security contributions prior to, or minor jobs parallel to, their non-registered activities. They may also have been registered as job seekers while being employed in one of these non-registered job types. Self-employed persons may also have been registered while receiving a business start-up allowance.

[6]For this reason, the very data used in this project cannot be made available to the scientific community. The data set most similar to the data used here is called "PASS survey data linked to administrative data of the IAB (PASS-ADIAB) 1975 - 2011"

3 Record linkage procedures

Before we describe the linkage procedures, it is important to acknowledge that there is no single method of performing a record linkage that can be universally applied in all applications. Record linkage is usually performed in multiple steps, and often subjective decision-making is needed to arrive at a final determination. Unfortunately, most survey-based linkage applications do not publish detailed descriptions of their linkage methods nor the decisions that were made during intermediate steps. The procedures that we use in this application closely resemble those used in other record linkages conducted at the IAB by the German Record Linkage Center (GRLC, see www.record-linkage.de for more details). In describing the procedures below, we have been as transparent as possible about the methods we used and the decisions we made.

3.1 Preprocessing

The first step of the linkage process was a thorough preprocessing of the aforementioned matching variables: *first name*, *last name*, *zip code*, *city*, *street name*, *house number*, *sex*, and dichotomized *birth cohort* (born before 1945 and after). This entailed consistently cleaning and standardizing the variables to make them comparable across the two databases.[7] The first steps involved substituting special characters and German umlauts as well as setting letters in uppercase. As the variable including the *first name* may contain more than one word, the variable was split into substrings using spaces as delimiters. As a result, four *first name* variables were created: one that contained all original parts of the *first name* written together in one word and up to three variables per person each containing one part of the *first name*. This allows for cases in which one of the data sources contains all of a person's given names (e.g. James Tiberius) while the other data source only contains one of them (e.g. only James or Tiberius). *Last names* were stored in one consecutive string.

The administrative linkage file potentially contains several records of personal data on a given person, in many cases even from different sources. We therefore retained all useful variation contained within different records on a given person (e.g. different number or spelling of given names, changed address, etc.). This allowed us to compare each version of a given identifier per person with the survey records, making it more likely that at least one of them was similar or equal to the identifier given in the respective other database. We also consolidated information across the records, if possible. Missing fields on *sex* and *birth year* were filled-in using valid information from other records on the same person. If different records on a given person showed conflicting information on his or her *sex*, we consolidated the records by deriving the most likely *sex* from a comprehensive list of given names and their most commonly associated sex. In case of conflicting birth years for a given person, either the modal or the most recent value was chosen, depending on whether an unambiguous mode was available.

3.2 Linkage steps

The actual linkage involved five consecutive steps: during the first step, we performed a stepwise deterministic linkage with two sub-steps; for the second and the third steps we used the software Merge ToolBox (MTB)[8] to conduct distance-based and probabilistic record linkage, respectively.[9] Our last two steps cover the process of discriminating real matches from false matches.

The first sub-step of the stepwise deterministic linkage used the variables *first name*, *last name*, *zip code*, *city*, *street name*, *house number*, *sex,* and *birth cohort* as matching variables. For *first name*, we only used the variable containing all given names written together in one word. The task of matching different *first name* combinations was left over to linkage steps using the MTB, which can handle this issue by an array-matching routine. In the second sub-step we used only seven matching variables, and only for persons that remained unmatched: *first name*, *last name*, *zip code*, *city*, *street name*, *sex*, and *birth cohort*.[10]

The remaining unmatched cases were the basis for the distance-based and probabilistic record linkage process. For the distance-based linkage we used the matching variables *first name*, *last name*, *zip code*, *city*, *street name*, *house number*, *sex*, and *birth cohort*. To increase efficiency of the comparison, we blocked by the first three digits of the *zip code* and by *sex*. That means we only compare records from both databases that show the same sub-*zip code* and the same *sex*, thereby immensely reducing the number of necessary comparisons.[11]

(2014). PASS-ADIAB contains survey data of linked consenters up to wave 5 as well as the administrative data described above. This data set can be accessed through the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the IAB. See Antoni and Bethmann (2014) for more details on the data and http://fdz.iab.de/en/FDZ_Individual_Data/PASS.aspx#ADIAB to request data access.

[7] The preprocessing routines have originally been developed by the GRLC. The same routines were used and adapted to this project.

[8] The MTB is freely available for academic purposes. For more details, see Schnell, Bachteler, and Bender (2004) or www.record-linkage.de

[9] See Gomatam, Carter, Ariet, and Mitchell (2002) for an empirical comparison of stepwise deterministic linkage with probabilistic linkage.

[10] The reason for the omission of *house number* in the second sub-step of the stepwise deterministic linkage is that this identifier shows an above average number of missings.

[11] For an overview of blocking methods, see Christen (2012, chapter 4).

For the variables *first name*, *last name*, *city*, and *street name* we applied bigrams[12] for the comparison. We did not calculate numerical distances for *zip code* or *house number*. For these identifiers, we instead checked whether they agreed completely. As our data contain more than one *first name* variable, we used array matching for that field. We therefore had the MTB compare all four *first name* variables from the survey sample frame with each of the eight possible *first name* variables in the administrative records. That way every available representation of this field in one database was compared with every available representation in the second database. Thereby we were able to retain the highest resulting similarity value from any of these comparisons.

For cases that could not be linked up to this point we subsequently used probabilistic record linkage see, Fellegi and Sunter (1969).[13] The linkage variables as well as the array matching procedure were the same as during the distance-based linkage step. We derived agreement weights from m- and u-probabilities[14] (e.g. Herzog, Scheuren, & Winkler, 2007, pp. 83-84) based on previous linkages of PASS data (see table 1). For all cases still remaining unmatched after this step we added a second probabilistic linkage step. To do so, we blocked only by *sex*. Thus, in contrast to previous steps, matches could now be found over all *zip code* regions.

3.3 Evaluation of deterministic linkage

During the linkage process we repeatedly took measures to discriminate true matches (true positives) from false matches (false positives). The evaluation step was to determine which of the multiple assignments resulting from the 1:m relation of the stepwise deterministic linkage represents the most adequate match. The challenge was that all of these matches were deterministic matches, and without further information it was not possible to decide which matched person should be chosen. In order to identify and solve multiple assignments we implemented a set of consolidation rules in an ordered sequence. The specific rules are provided in the appendix.

3.4 Evaluation of distance-based and probabilistic linkage

The overall challenge for the evaluation of distance-based and probabilistic linkage outcomes was to automatically apply consistent decision rules. In order to achieve that we developed a Stata tool that computes the Jaro-Winkler distance (see Winkler, 1990) for a set of identifiers (*first name*, *last name*, *street name*) as well as an indicator for agreement on *house number*. With this tool we were able to construct a match certainty index (MCI) to classify the best 17 possible matches offered by MTB within the respective linkage step. An overview of the MCI index values, identification rules, and their implementation is shown in table 2. The MCI values are ordered by decreasing strictness, or put differently, the higher the value, the higher the certainty that a true match has been identified. This table also shows that we extended the MCI to include cases that were matched during the stepwise deterministic linkage.[15]

The big advantage of the MCI compared to the original similarity measure is that it can be translated into specific and interpretable identification rules. From that we learn, for instance, that the difference between values 16 and 17 stems from the fact that the first deterministic linkage sub-step compares the *house number* of record pairs while the second sub-step does not. Each of the lower index values increasingly relaxes the requirement regarding the certainty, for example, by allowing the similarity on a given matching variable to be only 90 or 80 percent, respectively, instead of requiring full agreement.

[12] Bigrams are a special case of n-grams (or q-grams, depending on the terminology). N-grams are substrings of a string with the length *n*. A set of bigrams for the name MARY would be MA, AR and RY. In record linkage applications one most commonly encounters bigrams or trigrams (Stephen, 1994, p. 41).

[13] We put a strong emphasis on distance-based linkage by running it directly after the stepwise deterministic linkage and by only applying probabilistic linkage to cases that could not be linked with the first two methods. We acknowledge that it is possible that probabilistic linkage may have matched different record pairs, had it been applied before the distance-based step.

[14] The m-probability is the probability of two records agreeing on a given identifier when both records stem from the set of true matches. The u-probability is the probability of two records agreeing on a given identifier when the record pair stems from the set of true non-matches, i.e. the probability of two records from that set agreeing completely by chance. The ratio of these conditional probabilities quantifies how strongly an agreement on a given identifier indicates a match. For instance, the probability of agreement of a record pair on the identifier *sex* among the matches is 98.8%, whereas an agreement among the non-matches has a probability of 50%. This leads to an agreement weight of 1.976. Moreover, the number of possible values that an identifier can take on differs strongly between identifiers. While *sex* has only two possible values, the number is much higher for *first name* or *last name*. The probability of agreement on, for instance, the *last name* among the non-matches thus is only 0.05%. Depending on the comparison function, either an indicator for exact agreement (0/1) or the similarity derived from a comparison of bigrams enters the composite weight.

[15] Bear in mind that the identification rules shown in table 2 do not contain all identifiers that were used during the respective comparison step. For instance, although the identification rules 1 through 5 may use only very few identifiers to determine the MCI, the actual linkage steps still included the identifiers shown in table 1. The certainty of having linked the correct records thus is higher than the less restrictive identification rules may lead one to believe.

Table 1
M- and U-Probabilities and Comparison Function of Probabilistic Linkage Step

Matching Identifier	M-Probability	U-Probability	Comparison Function
Last name	.850	.0005	Bigrams
First name	.801	.0020	Bigrams
Sex	.988	.5000	Exact
Birth cohort	.900	.1000	Exact
Street name	.792	.0010	Bigrams
House number	.821	.0200	Exact
City	.876	.0120	Bigrams
Zip code	.889	.0490	Exact

Table 2
Overview of Match Certainty Index (MCI) Values, Identification Rules, and Linkage Steps

MCI Value	Applied Identification Rule [a,b]	Linkage Step
17	LN=1 & FN=1 & ST=1 & HN=1 & ZIP=1 & CI=1 & SEX=1 & BC=1	Stepwise deterministic linkage, sub-step 1
16	LN=1 & FN=1 & ST=1 & ZIP=1 & CI=1 & SEX=1 & BC=1	Stepwise deterministic linkage, sub-step 2
15	LN=1 & FN=1 & ST>0.8 & HN=1	Probabilistic linkage 1 & 2
14	LN=1 & FN=1 & ST>0.8 & HN<1	Probabilistic linkage 1 & 2
13	LN=1 & FN>0.9 & ST>0.8 & HN=1	Probabilistic linkage 1 & 2
12	LN=1 & FN>0.9 & ST>0.8 & HN<1	Probabilistic linkage 1 & 2
11	LN>0.9 & FN=1 & ST>0.8 & HN=1	Probabilistic linkage 1 & 2
10	LN>0.9 & FN=1 & ST>0.8 & HN<1	Probabilistic linkage 1 & 2
9	LN>0.9 & FN>0.9 & ST>0.8 & HN=1	Probabilistic linkage 1 & 2
8	LN>0.9 & FN>0.9 & ST>0.8 & HN<1	Probabilistic linkage 1 & 2
7	LN>0.8 & FN=1 & ST>0.8 & HN=1	Probabilistic linkage 1 & 2
6	LN=1 & FN=1 & ST>0.6 & HN=1	Probabilistic linkage 1 & 2
5	LN=1 & FN=1	Probabilistic linkage 1 & 2
4	LN=1 & ST=1 & HN=1	Probabilistic linkage 1
3	LN=1 & ST=1 & HN<1	Probabilistic linkage 1
2	LN=1	Distance-based linkage, Probabilistic linkage 1
1	LN>0.8 & FN>0.8	Distance-based linkage, Probabilistic linkage 1
0	None of the above	Not classified as match

[a] Abbreviations relate to the following matching variables: BC=*birth cohort*, CI=*city*, FN=*first name*, HN=*house number*, LN=*last name*, SEX=*sex*, ST=*street name*, ZIP=*zip code*.
[b] The numbers equated (or compared) to each matching variable represent the degree of similarity for the given variable, with 1 representing full agreement.

4 Results

4.1 Linkage rate by sample disposition codes

We report two linkage rates: an "inclusive" rate and a "restrictive" rate. The inclusive rate refers to all cases classified as a match (MCI values between 1 and 17; see table 2). The restrictive rate refers to all cases matched under a stricter matching criterion (MCI values between 6 and 17).[16] While the inclusive matches are likely less precise than the restrictive matches, we report both for comparison.

[16] When comparing characteristics collected from both survey and administrative data for the matched respondents who consented to linkage of their survey responses, we identified several false matches with a person from a different generation in the same household, mostly in the range of 20-30 years older than the sampled person (average age discrepancy: 3.69 years). This shows how important it would have been to receive more narrowly defined age groups instead of two broad birth cohorts as part of the sample frame information. To reduce the mismatches, we conducted a sen-

Both linkage rates are reported overall and by sample disposition codes as defined by AAPOR standards (AAPOR, 2016). The left part of Table 3 shows the inclusive linkage rate for each disposition code and the coefficients from a logistic regression model where the binary (inclusive) linkage outcome is regressed on all disposition codes. The right hand part presents the same results for the restrictive linkage.

Under the inclusive linkage criterion, a total of 5,015 (out of 6,237) persons/households are linked to the IAB administrative database; a linkage rate of 80.41 percent. There is some variability in linkage rates across disposition codes. The inclusive linkage rate is highest among noncontacts (86.65 percent), followed by unknown eligibility (85.40 percent), "other" nonresponse (82.30 percent), completed household interviews (81.17 percent), refusals (78.95 percent), not eligible (68.13 percent), and "not able" nonresponse (59.21 percent). The highest linkage rate among the noncontacts likely reflects people who are employed fulltime and are well-represented in the employment database, but are home less often making them difficult to reach. The lowest linkage rates among nonrespondents who were not eligible or not able to participate in the survey may reflect persons who are outside of the labor force and thus not well-represented in the administrative database.

Under the "restrictive" linkage criterion, a total of 3,734 persons/households are linked; a linkage rate of 59.87 percent. The restrictive linkage rate varies across sample disposition codes in nearly the same pattern (from highest to lowest) as the inclusive linkage rate: noncontacts (69.91 percent), unknown eligibility (67.57 percent), "other" nonresponse (64.44 percent), refusal (58.59 percent), completed household interview (57.47 percent), not eligible (42.86 percent), and "not able" nonresponse (42.11 percent).

4.2 Linkage rate by household characteristics

The remaining analyses examine how the linkage varied across household- and person-level characteristics that were collected in the survey. These analyses are possible without linkage consent because only the linkage indicator is merged to the survey responses.

First, we examine the following household characteristics collected from the household interview: composition (single-person, couple without children, couple or single parent with children, other), housing tenure (own, rent/other), material deprivation (index: 0,1,2,3+; Berg et al., 2012), net income (in Euros; <1,000, 1,000-1,999, 2,000-2,999, 3,000+, missing), savings (in Euros; <1,000, 1,000-10,000, 10,000-20,000, 20,000+), and whether or not Unemployment Benefit II (UB II) was ever received since 2009. These variables were selected without any prior expectations; however, we do expect linkage rates to be higher for households who had ever received UB II as the BA is responsible for administering this benefit and it is automatically recorded in the administrative database.

The left part of Table 4 shows the inclusive linkage rate for each household characteristic and the coefficients from a logistic regression model where the binary (inclusive) linkage outcome is regressed on all household variables. The right hand part presents the same results for the restrictive linkage. Both linkages reveal similar results: linkage rates are lower for households without children, household ownership, households with high material deprivation, and households with savings of at least 20,000 EUR. As expected, we find that households that have ever received UB II since 2009 are linked at a higher rate than households that have not received this benefit. Linkage rates did not consistently differ by net income categories.

4.3 Linkage rate by person-level characteristics

The last set of analyses reviews how the linkage rate varied across the following person-level characteristics collected from the personal interview with the sampled person: age (in years; <40, 40-54, 55-69, 70+), sex, marital tatus (never married/partnered, ever married/partnered), type of postsecondary education (vocational training, college degree, none/other), foreign born, type of health insurance (social, private/other), and employment status (employed-blue collar, employed-white collar, employed-other/self-/civil servant, not employed).

We have expectations for some of these variables. Self-employed persons and civil servants are likely to be linked at lower rates than other employed groups because they do not make social security contributions and thus are not covered in the administrative database.[17] Second, we expect people with private health insurance to be linked at a lower rate than people with social health insurance. The reason being that privately insured individuals tend to be better educated and more successful in the labor market – if you are not a civil servant or self-employed, you can only be privately insured if your earnings are above a certain threshold. Therefore we expect that these people show up in the job search or unemployment data much more rarely than those in social health insurance. Lastly, we expect that older persons will be linked at lower rates than younger persons for two reasons: 1) the older the person, the more likely they will have left the labor force completely (the mandatory retirement age in Germany was 65 at the time of PASS wave 5) and those who left the

sitivity analysis by cumulatively removing the least certain matches (starting with MCI = 1, 2, and so on) until a reasonable balance was struck between reducing the overall age discrepancy and retaining as many true matches as possible (MCI values between 6 and 17). This "restrictive" match criterion reduced the average age discrepancy to 0.34 years.

[17] As mentioned before, people in these groups may still show up in the administrative data, e.g. if they previously held jobs that were subject to social security contributions.

Table 3
Linkage Rates and Logistic Regression Coefficients of Linkage Outcome (1=Link; 0=Non-Link) by Household Sample Disposition Codes.[a]

		Inclusive linkage					Restrictive linkage				
				Logistic Regression					Logistic Regression		
Household Sample Disposition Codes	N	Linkage Rate	N Linked	Coef.	S.E.	p-value	Linkage Rate	N Linked	Coef.	S.E.	p-value
Unknown eligibility	404	85.40	345	0.31	0.21	.149	67.57	273	0.43	0.15	.005
Not eligible	91	68.13	62	−0.70	0.26	.009	42.86	39	−0.59	0.25	.025
NR-no contact	442	86.65	383	0.41	0.25	.108	69.91	309	0.54	0.18	.004
NR-not able	76	59.21	45	−1.09	0.31	.001	42.11	32	−0.62	0.32	.058
NR-refusal	3,040	78.95	2,400	−0.14	0.09	.113	58.59	1,781	0.05	0.07	.523
NR-other	644	82.30	530	0.08	0.16	.634	64.44	415	0.29	0.10	.007
Interviewed	1,540	81.17	1,250	–	–	–	57.47	885	–	–	–
Total / Intercept	6,237	80.41	5,015	1.46	0.09	.000	59.87	3,734	0.30	0.06	.000

[a] Point estimates and their standard errors (S.E.) are adjusted for all complex sample design features.

Table 4
Linkage Rates and Logistic Regression Coefficients of Linkage Outcome (1=Link; 0=Non-Link) by Household-Level Characteristics Collected from Interview with Household Respondent[a]

		Inclusive linkage					Restrictive linkage				
				Logistic Regression					Logistic Regression		
Household Level Characteristics	N	Linkage Rate	N Linked	Coef.	S.E.	p-value	Linkage Rate	N Linked	Coef.	S.E.	p-value
Household composition											
Single-person household	316	80.05	249	–	–	–	60.54	182	–	–	–
Couple w/o children	577	72.74	417	−0.44	0.26	.095	47.48	265	−0.18	0.26	.506
Coup./sing. parent w/ children	524	91.69	477	0.78	0.48	.111	70.97	351	0.69	0.35	.057
Other	93	93.04	87	1.16	0.98	.243	81.84	72	1.49	0.48	.003
Housing tenure											
Own	791	75.30	620	−0.98	0.27	.001	45.99	375	−1.18	0.17	.000
Rent/other	719	85.66	610	–	–	–	70.21	495	–	–	–
Material deprivation index											
0 (least deprived)	851	85.06	729	–	–	–	61.21	509	–	–	–
1	328	83.86	267	−0.18	0.27	.495	60.45	188	−0.39	0.22	.087
2	192	74.60	133	−1.06	0.28	.000	61.95	99	−0.74	0.28	.010
3+ (most deprived)	139	73.04	101	−1.42	0.32	.000	52.38	74	−1.54	0.32	.000
Net income in past month											
<1,000 EUR	130	84.65	105	–	–	–	74.35	89	–	–	–
1,000–1,999	360	78.70	283	−0.25	0.44	.568	62.13	217	−0.48	0.37	.199
2,000–2,999	330	83.63	269	0.10	0.54	.855	56.31	184	−0.65	0.50	.204
3,000+	448	83.78	380	0.08	0.62	.895	56.51	247	−0.54	0.46	.247
Missing	242	76.20	193	−0.51	0.50	.313	51.56	133	−0.70	0.43	.108

Continues on next page

Continued from last page

		Inclusive linkage					Restrictive linkage				
				Logistic Regression					Logistic Regression		
Household Level Characteristics	N	Linkage Rate	N Linked	Coef.	S.E.	p-value	Linkage Rate	N Linked	Coef.	S.E.	p-value
Saving amount											
<1,000 EUR	394	84.39	331	–	–	–	71.32	266	–	–	–
1,000–10,000	406	83.27	343	0.13	0.24	.600	62.88	256	−0.05	0.23	.836
10,000–20,000	172	82.85	138	0.09	0.26	.736	60.97	97	−0.02	0.26	.951
20,000+	359	73.68	271	−0.44	0.31	.164	44.26	161	−0.68	0.25	.009
Missing	179	80.89	147	0.27	0.41	.506	50.21	90	−0.36	0.27	.196
Received UB II since 2009											
Yes	123	97.02	119	2.14	1.01	.041	93.73	114	2.12	0.67	.003
No	1,385	79.27	1,109	–	–	–	55.69	755	–	–	–
Total / Intercept	1,510[b]	81.30	1,230	2.41	0.47	.000	60.01	870	1.76	0.45	.000

[a] Point estimates and their standard errors (S.E.) are weighted and adjusted for complex sample design features.

[b] Household-level interviews were completed with 1,540 households, but no person-level interviews were conducted with 30 of these households. PASS does not release survey data for households in which no person-level interviews are conducted, which is why the base total for this table is 1,510 households.

labor force completely prior to 2009 (the starting year that addresses were extracted for the linkage) would not be covered by the linkage identifiers from the administrative data; and 2) older persons are more likely to have been employed as civil servants because this status was more common in the past and provided a very high level of job security that resulted in very stable employment relationships.

The left part of Table 5 presents the inclusive linkage rate for each person-level variable and the coefficients from a logistic regression model where the (inclusive) binary linkage outcome is regressed on all person-level variables. The right hand side of the table presents the same figures under the restrictive linkage. The inclusive and restrictive linkage results yield generally similar patterns. As expected, older persons are linked at a substantially lower rate than younger persons. For example, only 8 percent of persons aged 70 and older are linked under the restrictive criterion. This is in contrast to a 40 percent linkage rate under the inclusive criterion for this group, which likely reflects some generational mismatches that were found during the linkage evaluation (as explained in an earlier footnote). Also in line with our expectations: civil servants and the self-employed are linked at lower rates than other employed groups, and privately insured persons are linked at a lower rate than those who have social health insurance. In the regression models, the likelihood of a link did not significantly differ by foreign birthplace,[18] marital status, and postsecondary education, but did differ by sex: males were more likely to be linked than females.

5 Discussion

In this case study we examined the feasibility of indirectly linking federal administrative records to a general population sample of survey respondents and nonrespondents. This study is particularly relevant at a time when auxiliary data sources are being sought to address survey methodological research inquiries. This particular application of linking administrative data to a nationally-representative sample of respondents and nonrespondents is uncommon in the sense that the sample was not originally drawn from the target administrative database and no unique identifier was available to perform a direct linkage. However, the lack of an administrative sampling frame and unique identifier is representative of most survey settings outside of certain countries that draw their samples from administrative sources (e.g. the Netherlands, Sweden); thus, we believe that our case study can be informative to other surveys considering similar sample linkages.

The study yielded four main findings. First, about 60 percent of individuals sampled in the survey could be linked to an administrative record under a strict linkage criterion; 80 percent of the sample could be linked under a more relaxed linkage criterion. Second, the distribution of linkages differed across most sample disposition codes and nonresponse types, with non-contacts linked at a slightly higher rate than refusals and interviewed households. Third, among the responding households, linkage rates varied by several household characteristics: presence of children, household ownership, material deprivation, household savings, and Unemployment Benefit II receipt, but not by net income. Lastly, among the interviewed respondents, linkage rates varied by some person-level characteristics, including age, sex, and type of employment and health insurance, but did not significantly vary by others, including marital status, postsecondary education, and foreign birth status.

It is unrealistic to expect that all sampled units drawn from the general population can be linked to any particular administrative source, especially one that is not intended to completely cover the general population, as was the case here and in other similar linkage applications (Bee et al., 2015). The selectivity of the linkage raises several important questions concerning the utility of the linked administrative records. For instance, is it appropriate to use incomplete, non-randomly linked administrative data to support survey methodological activities? If not, can the administrative data for the non-linked cases be imputed or adjusted statistically to minimize linkage bias? Do linkage errors interact with other survey errors (e.g. coverage, nonresponse)? All of these issues warrant further research to obtain a better understanding of the utility of these linked data.

Provided these issues can be resolved, we see several opportunities afforded by linking survey samples to federal administrative records. First, given that federal administrative databases – including the one used here – often contain substantive variables that are likely associated with key survey topics (e.g. income, benefit receipt, employment history) these administrative variables could prove to be particularly useful for evaluating survey nonresponse and measurement error. The longitudinal nature of these data also affords the possibility of studying these error sources over time. The administrative data could be useful for explaining why individuals do not participate in surveys over multiple waves (e.g. due to a location or employment change), and whether their non-participation is associated with administrative variables that are likely to be related to key survey variables. We will be monitoring this closely in the coming waves of the

[18] Had there been a different likelihood of linkage for people born abroad we would have to assume a different or additional reason for why people may not be found during the linkage. So far, characteristics showing strong correlations with linkage success, e.g. the employment status, are probably related to the likelihood of being in the administrative data at all. Being born abroad, on the other hand, may also be related to having les common names. That makes spelling mistakes or inconsistencies more likely. These in turn make it more likely that the identifying information within the two data sources do not match, even though the person may actually be in both data sources. However, it is impossible to distinguish between these two possible reasons for not being linked.

Table 5
Linkage Rates and Logistic Regression Coefficients of Linkage Outcome (1=Link; 0=Non-Link) by Person-Level Characteristics Collected from Interview with Sampled Person[a]

		Inclusive linkage					Restrictive linkage				
				Logistic Regression					Logistic Regression		
Person-Level Characteristics	N	Linkage Rate	N Linked	Coef.	S.E.	p-value	Linkage Rate	N Linked	Coef.	S.E.	p-value
Age (in years)											
<40	316	96.81	308	–		–	84.64	278	–		–
40–54	460	92.71	421	−0.87	0.85	.314	73.40	307	−0.55	0.32	.092
55–69	388	81.23	322	−1.74	0.68	.014	55.38	218	−1.44	0.31	.000
70+	276	40.05	122	−3.63	0.61	.000	7.92	27	−4.36	0.34	.000
Sex											
Female	705	75.44	556	−0.81	0.21	.001	53.48	385	−0.75	0.22	.002
Male	734	87.13	616	–	–	–	67.55	444	–	–	–
Marital status											
Never married/partnered	308	95.93	292	0.40	0.64	.536	85.06	255	0.62	0.38	.114
Ever married/partnered	1,120	77.30	871	–	–	–	53.89	568	–	–	–
Postsecondary education											
None/other	254	84.36	210	–	–	–	63.22	162	–	–	–
Vocational training	890	80.43	737	−0.08	0.35	.821	61.50	530	0.47	0.34	.170
College degree	296	79.49	226	0.01	0.40	.990	52.26	138	0.42	0.35	.234
Foreign born											
Yes	168	87.05	140	−0.49	0.51	.344	68.84	109	−0.18	0.36	.608
No	1,270	79.90	1,031	–	–	–	58.61	720	–	–	–

Continues on next page

Continued from last page

Person-Level Characteristics	N	Inclusive linkage					Restrictive linkage				
		Linkage Rate	N Linked	Logistic Regression			Linkage Rate	N Linked	Logistic Regression		
				Coef.	S.E.	p-value			Coef.	S.E.	p-value
Type of health insurance											
Private/other	217	75.48	161	−0.77	0.27	.006	38.86	80	−1.34	0.30	.000
Social	1,220	82.01	1,009	–	–	–	63.66	749	–	–	–
Employed											
Blue collar	124	96.69	120	–	–	–	78.64	99	–	–	–
White collar	410	95.93	389	−0.05	0.61	.932	82.08	311	0.43	0.41	.297
Self-employed/ civil servant/other	132	81.17	107	−1.29	0.47	.009	38.72	46	−0.91	0.55	.105
Not employed	770	71.32	554	−0.80	0.32	.018	49.56	372	0.33	0.50	.515
Total / Intercept	1,440[b]	81.21	1,173	4.64	0.52	.000	60.42	830	1.66	0.66	.016

[a] Point estimates and their standard errors (SE) are weighted and adjusted for complex sample design features.

[b] As noted in table 4, there were 1,510 households for which a household-level interview and a person-level interview (with at least the sampled person) was conducted. However, for 70 person-level interviews, we could not identify the sampled person on the household roster and thus could not retrieve their responses. This means that either the interviewer selected a household he/she was not supposed to interview, or that the sampled person used an alias which they are allowed to do if they have privacy concerns about their real name. But usually these aliases are names like "oldest child" or "partner" and this is not very frequent. For this reason, we only examine the person-level survey data for the 1,440 sampled individuals who were clearly recognized on the household roster.

PASS survey as the refreshment sample continues to mature. Lastly, assuming the administrative data can be linked prior to survey data collection, these data could be used in responsive designs to inform indicators of sample representativeness and facilitate targeted recruitment of specific subgroups.

It is important to acknowledge some idiosyncratic features of this case study. First, the administrative employment database considered here is not ideal for general population linkages as it is not intended to cover the entire population, nor the entire working population. Second, as with all indirect linkage methods, there is a chance that some matches are false. We investigated the extent of this issue using the survey data for the consenting respondents and adjusted the linkage criterion accordingly, but we concede that the probability of any remaining mismatches is still greater than zero. Third, the PASS refreshment sample was drawn directly from municipality registration offices and included names, addresses, and other pertinent details that facilitated the linkage. We acknowledge that many general population surveys do not have access to such detailed information. Though with the rise in commercial and address-searchable databases (Smith & Kim, 2013), we speculate such information may become more accessible in the future.

An important consideration not addressed in this case study is the ethical issues associated with unconsented access and linkage of administrative data to survey paradata. Because all data sources used and generated for this project belong to the IAB and because there was no need to link actual survey responses to administrative records, the legal authorities deemed it unnecessary to obtain consent from responding and nonresponding individuals/households for purposes of this project. We acknowledge that this situation may not apply to other research settings, particularly when multiple independent parties are involved, including the survey sponsor, the data collection agency, and the administrative data owners. In particular, legal and ethical issues arise when linking information to nonrespondents. Sakshaug and Eckman (in press) demonstrate that obtaining consent from survey nonrespondents to use their administrative records is feasible, but the consent rate they report is significantly lower than the unconsented linkage rate achieved in the present study; a result that further brings the selectivity of linkage into question. Ultimately, each relevant data party has to work with the appropriate authorities to discuss the principles of conducting this type of research and to what extent consent may or may not be needed.

While much has been discussed about appending survey samples with various sources of auxiliary data, we hope our study brings more attention to federal administrative data sources in these discussions. We do not claim that these data will be an effective source of auxiliary data in all survey applications, as many technical and logistical issues exist, but we do advocate further research into ways in which these data could be more broadly linked to general population samples and leveraged to improve surveys and the quality of data they produce. We encourage survey organizations, particularly those that already link survey responses to federal administrative databases, to have this conversation with the appropriate administrative data authorities and research ethics committees.

References

Antoni, M. & Bethmann, A. (2014). *PASS-Befragungsdaten verknüpft mit administrativen Daten des IAB (PASS-ADIAB) 1975-2011* (FDZ-Datenreport No. 03/2014). Institute for Employment Research, Nuremberg.

Antoni, M., Ganzer, A., & vom Berge, P. (2016). *Sample of integrated labour market biographies (SIAB) 1975-2014* (FDZ-Datenreport No. 04/2016 (en)). Institute for Employment Research, Nuremberg.

Antoni, M. & Seth, S. (2012). ALWA-ADIAB-linked individual survey and administrative data for substantive and methodological research. *Schmollers Jahrbuch*, *132*(1), 141–146.

Bee, C. A., Gathright, G., & Meyer, B. D. (2015). *Bias from unit non-response in the measurement of income in household surveys*. Paper presented at the Joint Statistical Meetings of the American Statistical Association. Retrieved from http://www.sole-jole.org/16068.pdf

Berg, M., Cramer, R., Dickmann, C., Gilberg, R., Jesske, B., Kleudgen, M., … Wurdack, A. (2012). *Codebook and documentation of the panel study 'Labour Market and Social Security' (PASS) Wave 5* (FDZ-Datenreport No. 06/2012 (en)). Institute for Employment Research, Nuremberg.

Blom, E. & Carlsson, F. (1999). *Registers in official statistics: a Swedish perspective*. Invited paper for the Joint/ECE/Eurostat Work Session on Registers and Administrative Records for Social and Demographic Statistics, Geneva. Retrieved from http://www.unece.org / fileadmin / DAM / stats / documents / 1999 / 03 / registers/14.e.pdf

Chetty, R. (2012). *Time trends in the use of administrative data for empirical research*. Presented at the National Bureau of Economic Research, Summer Institute, Cambridge, Mass. Retrieved from ttp : // www . rajchetty.com/chettyfiles/admin_data_trends.pdf

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

Couper, M. P. (1998, August). *Measuring survey quality in a CASIC environment*. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, TX. Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf

Couper, M. P. & Lyberg, L. E. (2005, April). *The use of paradata in survey research*. Paper presented at the 54th Session of the International Statistical Institute, Sydney, Australia.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

Diez Roux, A. V. (2001). Investigating neighborhood and area effects on health. *American Journal of Public Health*, *91*(11), 1783–1789.

DiSogra, C., Dennis, J. M., & Fahimi, M. (2010, August). *On the quality of ancillary data available for address-based sampling*. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Vancouver, British Columbia. Retrieved from http://www.knowledgenetworks.com/ganp/docs/jsm2010/On-the-Quality-of-Ancillary-ABS-2010-JSM-submission.pdf

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

Freedman, V. A., McGonagle, K., & Andreski, P. (2014). *The panel study of income dynamics linked medicare claims data* (PSID Technical Report Series No. 14-01). University of Michigan.

Gomatam, S., Carter, R., Ariet, M., & Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, *21*(10), 1485–1496.

Groves, R. M. & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457.

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

Jacobebbinghaus, P. & Seth, S. (2007). The German integrated employment biographies sample IEBS. *Schmollers Jahrbuch*, *127*(2), 335–342.

Knies, G. & 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*.

Korbmacher, J. & Czaplicki, C. (2013). Linking SHARE survey data with administrative records: First experiences from SHARE-Germany. In F. Malter & A. Börsch-Supan (Eds.), *Hare wave 4: innovations & methodology* (pp. 47–53). Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information*. Hoboken, New Jersey: Wiley.

Möller, J. & Walwei, U. (2009). Editorial. In S. Koch, P. Kupka, & J. Steinke (Eds.), *Aktivierung, Erwerbstätigkeit und Teilhabe. Vier Jahre Grundsicherung für Arbeitsuchende* (pp. 11–12). Bielefeld: IAB-Bibliothek 315.

Olson, J. A. (1999). Linkages with data from Social Security administrative records in the Health and Retirement Study. *Social Security Bulletin*, *62*(2), 73–85.

Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science*, *645*(1), 142–170.

Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., & DiSogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly*, *78*(4), 889–916.

PASS survey data linked to administrative data of the IAB (PASS-ADIAB) 1975-2011. (2014). Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

Peytchev, A. & Olson, K. (2007, August). *Using interviewer observations to improve nonresponse adjustments: NES 2004*. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Salt Lake City, UT. Retrieved from https://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000695.pdf

Raghunathan, T. E. & Hoewyk, J. V. (2008). *Disclosure risk assessment for survey microdata*. Unpublished Manuscript, University of Michigan.

Sakshaug, J. W. & Eckman, S. (in press). Are survey nonrespondents willing to provide consent to use administrative records? Evidence from a nonresponse follow-up survey in Germany. *Public Opinion Quarterly*.

Schnell, R., Bachteler, T., & Bender, S. (2004). A toolbox for record linkage. *Austrian Journal of Statistics*, *33*(1–2), 125–133.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, *27*(2), 1–24.

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations? *Public Opinion Quarterly*, *78*(2), 440–473.

Smith, T. W. & Kim, J. (2013). An assessment of the multilevel integrated database approach. *The Annals of the American Academy of Political and Social Science*, *645*(1), 185–221.

Stephen, G. A. (1994). *String searching algorithms*. Singapore: World Scientific.

The American Association for Public Opinion Research (AAPOR). (2016). *Standard definitions: Final dispo-*

sitions of case codes and outcome rates for surveys (9th ed.). AAPOR. Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Trappmann, M., Beste, J., Bethmann, A., & Müller, G. (2013). The pass panel survey after six waves. *Journal for Labour Market Research*, *46*(4), 275–281.

United Nations Economic Commission for Europe (UNECE). (2007). *Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics* (Technical Report No. E.07.II.E.11). Retrieved from http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

Vannieuwenhuyze, J. T. A. & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, *42*(1), 82–104.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, *76*(3), 555–575.

Wallgren, A. & Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*. West Sussex, England: Wiley.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 211–225.

West, B. T., Kreuter, F., & Trappmann, M. (2014). Is the collection of interviewer observations worthwhile in an Economic Panel Survey? New evidence from the German Labor Market and Social Security (PASS) study. *Journal of Survey Statistics and Methodology*, *2*(2), 159–181.

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The utility of alternative commercial data sources for survey operations and estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, *3*(2), 240–264.

Winkler, W. E. (1990, August). *String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage*. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Anaheim, CA. Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf

Appendix

Consolidation rules for multiple assignments

Rule 1: consider an assignment to be valid if the date of last contact (information from paradata) lies in the time-frame of the administrative record. For non-contacted persons, the contact date was set on the mean value of contacted persons.

Rule 2: if the end-date of the time-frame is closer to the contact date than with the other assignments, then consider this assignment to be valid.

Rule 3: if the begin-date of the time-frame is closer to the contact date than with the other assignments with the same distance of end-date to contact date, then consider this assignment to be valid.

Rule 4: if previous rules do not solve multiple assignments, consider the currentness of the administrative record. Classify the most recently created record to be valid. If a given rule did not solve the case of multiple assignments, the case was moved to the next rule, and so on.