DE GRUYTER OPEN

Journal of Official Statistics, Vol. 34, No. 4, 2018, pp. 909–933, http://dx.doi.org/10.2478/JOS-2018-0045

Statistical Matching as a Supplement to Record Linkage: A Valuable Method to Tackle Nonconsent Bias?

Jonathan Gessendorfer[1], *Jonas Beste*[2], *Jörg Drechsler*[2], *and Joseph W. Sakshaug*[2]

Record linkage has become an important tool for increasing research opportunities in the social sciences. Surveys that perform record linkage to administrative records are often required to obtain informed consent from respondents prior to linkage. A major concern is that nonconsent could introduce biases in analyses based on the linked data. One straightforward strategy to overcome the missing data problem created by nonconsent is to match nonconsenters with statistically similar units in the target administrative database. To assess the effectiveness of statistical matching in this context, we use data from two German panel surveys that have been linked to an administrative database of the German Federal Employment Agency. We evaluate the statistical matching procedure under various artificial nonconsent scenarios and show that the method can be effective in reducing nonconsent biases in marginal distributions, but that biases in multivariate estimates can sometimes be worsened. We discuss the implications of these findings for survey practice and elaborate on some of the practical challenges of implementing the statistical matching procedure in the context of linkage nonconsent. The developed simulation design can act as a roadmap for other statistical agencies considering the proposed approach for their data.

Key words: Data fusion; survey data; administrative data; linkage nonconsent.

1. Introduction

Many survey organizations link their surveys to large-scale administrative databases in order to increase research opportunities, minimize data collection costs, and enhance data utility (Calderwood and Lessof 2009). To give only a few examples, the Avon Longitudinal Study of Parents and Children (Ness 2004) and the UK Millennium Cohort Study (Mostafa 2016) link interview data to various health and social administrative records. Statistics Netherlands conducts linkages of surveys and various administrative registers to conduct the Dutch census (Schulte Nordholt et al. 2014). In Germany, several

[1] Email: jonathan.gessendorfer@gmail.com
[2] Institute for Employment Research, Regensburger Str. 100, 90478 Nuremberg, Germany. Emails: jonas.beste@iab.de, joerg.drechsler@iab.de, and joe.sakshaug@iab.de

Acknowledgments: We thank the anonymous reviewers, the associate editor and the guest editor, whose insightful comments and suggestions helped improve the manuscript considerably. This article uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi:10.5157/NEPS:SC6:7.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. This study also uses the factually anonymous data of the Panel Study 'Labour Market and Social Security' (PASS) and the factually anonymous Sample of Integrated Labour Market Biographies (SIAB). For both, data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

surveys, including the study "Working and Learning in a Changing World" (ALWA; Antoni and Seth (2011)) and the Socio-Economic Panel (SOEP) Migration Sample (Brücker et al. 2014), link to the Integrated Employment Biographies (IEB) – an administrative database of the German Federal Employment Agency (BA) that covers nearly the entire German population of employable age (Jacobebbinghaus and Seth 2010).

Due to data protection regulations, surveys in many countries are required to obtain informed consent from respondents prior to record linkage. In the European Union, for example, this requirement is part of the General Data Protection Regulation (GDPR 2016). Nonconsent and other reasons for record linkage failure lead to incomplete data and therefore a reduction in statistical power and precision of statistical estimates. Reviews of the linkage consent literature show that the amount of incomplete data can be quite severe with linkage consent rates below 50% in several studies (Sakshaug and Kreuter, 2012; da Silva et al. 2012). Even more alarming is the fact that linkage consent rates have been declining over time (Fulton 2012). Given this declining trend and low observed consent rates, there is increasing concern that nonconsenters could be systematically different from consenters, introducing bias in subsequent analyses based on the linked data. Numerous studies have demonstrated the biasing effects of nonconsent in actual linkage applications (Jenkins et al. 2006; Sakshaug and Huber 2016; Sakshaug et al. 2012; Sala et al. 2012; Mostafa 2016). Some of the most common variables affected by nonconsent bias include socio-demographics (for example, age, sex, ethnicity), economic variables (for example, income, income assistance benefits), and socio-environmental variables (for example, urbanicity, regional variation). While the majority of such biases have been found in survey variables, biases in the linked administrative variables have also been identified (Sakshaug and Kreuter 2012; Sakshaug and Vicari 2017; Sakshaug et al. 2017), suggesting that neither data source is immune to nonconsent bias.

Nonconsent generates a very specific missing data situation. In many ways, it is similar to the situation created by unit nonresponse if auxiliary information is available for both respondents and nonrespondents. However, an important difference is that the amount of information available for both consenters and nonconsenters — the data obtained from the survey — typically far exceeds the amount of information available for both respondents and nonrespondents. It is not obvious whether best practice methods developed to reduce nonresponse biases (Brick and Kalton 1996) would perform similarly for the missing data situation generated by nonconsent to record linkage. While extensive research has been done on optimizing linkage consent rates at the design stage – for example, by improving the wording or placement of the consent question in the survey (Kreuter et al. 2016) – no general guidelines have been proposed to reduce linkage nonconsent bias post-survey data collection.

One strategy to overcome the missing data problem induced by linkage nonconsent is to use statistical matching (Rässler 2002; D'Orazio et al. 2006b). Statistical matching methods merge individual records from two (or more) data sources based on their similarity on variables observed in all data sources. The main goal of the research presented here is to investigate the idea of performing statistical matching on the nonconsenting cases in order to 1) make the administrative data available for all survey participants, including linkage nonconsenters; and 2) reduce linkage nonconsent biases in estimates derived from the linked survey and administrative data. We evaluate this

strategy through a case study involving two major surveys in Germany that link to the IEB database: the "National Educational Panel Study" (NEPS) and the Panel "Labour Market and Social Security" (PASS).

The remainder of this article is organized as follows. In Section 2 we review record linkage and statistical matching as tools for combining information from different sources. In Section 3 we illustrate how statistical matching may be used as a supplement to record linkage for nonconsent bias reduction. Section 4 discusses problems in practice including why extensions to the classical statistical matching approaches, although promising from a methodological perspective, cannot be used in this context. In Section 5 we describe the two surveys and the administrative data source used to evaluate the proposed methodology.

In Section 6 we describe the study design and evaluation procedures. The results of the evaluation are presented in Section 7. The article concludes with a discussion of the case study results, their implications for survey practice, and practical issues associated with implementing the proposed methodology.

2. Record Linkage and Statistical Matching

To facilitate our review of record linkage and statistical matching we introduce the following notation. Let A and B be two data sets to be merged where vectors of random variables (X, Y) are observed in data set A and vectors of random variables (X, Z) are observed in data set B. For brevity, we limit our discussion to the most common scenario of merging two data sets. The goal of both record linkage and statistical matching is to use the information from both data sets in order to estimate the joint density $f(x, y, z)$ of the combined vector (X, Y, Z) in the population.

2.1. Record Linkage and Reasons for Unsuccessful Linkage

Record linkage techniques aim to identify and merge records of different data sources that refer to the same entity (Herzog et al. 2007). In the context of survey and administrative data linkage, the goal is to identify administrative data records that belong to the same survey respondents. Assuming the survey respondents represent a random subset of the population and linkage is successful for every respondent, the merged vectors consist of random realizations of (X, Y, Z) and thus inference regarding $f(x, y, z)$ is straightforward.

However, there are many reasons why record linkage may be unsuccessful for some survey respondents. One of those reasons is that record linkage techniques sometimes fail to identify true links. This happens particularly if only imperfect linkage identifiers – such as name and address information, instead of unique identifiers like national identification numbers – are used to merge both data sets. Fellegi and Sunter (1969) provide a mathematical framework for this situation. Imperfect identifiers can produce nonlinks or false positive links – merging of records that do not belong to the same unit – which can lead to attenuated associations between Y and Z. Another reason for unsuccessful record linkage is that some survey respondents might not have records in the administrative data set. If individuals with specific traits are missing systematically from the administrative data, this can also lead to biased inferences.

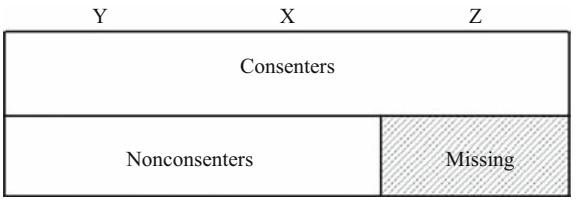

Fig. 1. The missing data situation in the combined data set.

A further reason for unsuccessful linkage – and the focus of this article – is due to the fact that data protection regulations often require that informed consent be obtained from survey respondents prior to data linkage. This creates a missing data situation in the combined data set as depicted in Figure 1. As noted in the introduction, contemporary research shows that linkage nonconsent is prevalent in surveys and can introduce bias in linked survey and administrative variables if only the completely observed parts of the data are used for analyses. Thus, methods to mitigate nonconsent bias are needed to obtain valid inferences from linked data sets.

Several methods have been developed and evaluated that deal with a very similar problem: unit nonresponse bias in surveys. The two most prominent methods are weighting adjustments and (multiple) imputation. Both are applicable to the linkage nonconsent scenario but have significant drawbacks in this context. The main drawback of weighting for linkage nonconsent is that analyses can only be performed on the consenting cases. That is, after constructing the weights, the survey information for all nonconsenters is completely ignored, making the approach inefficient especially for high nonconsent rates.

If the goal is to obtain a complete, rectangular data set on (X, Y, Z), multiple imputation can be considered to fill in the missing linked data. However, most imputation routines need good parametric models for the missing Z variables. The modeling step can become highly complex in the context of merging survey and administrative data, as the structure of administrative data is often not suitable for parametric specification. For example, administrative variables from the IEB database are measured in terms of spells with varying beginning and endpoints (Jacobebbinghaus and Seth 2010). Creating good parametric models for such variables is a very difficult and labor-intensive task.

Nearest neighbor hot-deck imputation is a possible alternative to parametric imputation (Chen and Shao, 2000; Andridge and Little 2010). This method may be used to identify a consenting respondent who is similar in (X, Y) to a nonconsenting respondent.

The consenting respondent then donates the observed Z to the observed X and Y data of the nonconsenting respondent. However, the feasibility of nearest neighbor hot-deck imputation depends heavily on the consent rate and on the sample size of the survey, since hot-deck runs into problems if the donor pool is sparse (Andridge and Little 2010).

2.2. *Review of Statistical Matching*

Statistical matching, sometimes known as data fusion, aims to integrate multiple data sources to draw inference on $f(X, Y, Z)$. Micro approaches to statistical matching create a synthetic data set, where X, Y and Z are available as if they were jointly observed, whereas

macro approaches attempt to draw inference on parameters that are nonestimable using only the separate data sets. We only consider micro approaches here, as macro approaches are less suited for the application we are considering in this article. For an overview of macro approaches, see D'Orazio et al. (2006b). This reference is also recommended for further information on the micro approaches we discuss below.

In standard statistical matching applications, data sets A and B are both random samples drawn from a much larger population. In this scenario, record linkage would be infeasible, as there is unlikely to be any overlap between the two data sources. Traditional approaches to statistical matching use a set of common variables X to combine A and B. For example, nearest neighbor matching techniques merge data of units that are similar in X. Specifically, for each unit in A – the recipients – a unit in B that is similar in X donates its Z information to the observed (X, Y) vector of the recipient (see Figure 2 for a visualization). Besides nearest neighbor, there are various other traditional statistical matching techniques. Beyond nonparametric methods, like nearest neighbor, fully parametric models or mixtures of parametric models and nonparametric matching techniques have been suggested in the literature (Rässler 2002; D'Orazio et al. 2006b).

Given that only the information in X is used in all traditional matching procedures, the distribution of (X, Y, Z) after statistical matching $\tilde{f}(x, y, z)$ will necessarily have a very specific characteristic: conditional on X, Z and Y will be independent:

$$\tilde{f}(y|x,z) = \tilde{f}(y|x) \quad \wedge \quad \tilde{f}(z|x,y) = \tilde{f}(z|x) \tag{1}$$

Therefore, if the aim is to draw inference regarding the relationship of Y and Z, one must implicitly assume that the two variables are independent conditional on X in the population. This is referred to as the conditional independence assumption (CIA). If the assumption is not met, the joint distribution of (X, Y, Z) after statistical matching will differ from the true distribution. Potentially, this can lead to biased inferences from the statistically matched data set (Sims 1972; Rodgers 1984; D'Orazio et al. 2006b). For example, correlations between Y and Z variables will typically be biased towards zero, as only the part of the correlation that can be explained by the X variables will be preserved in the statistically matched file. Similar to the effect of omitted variables, regression coefficients in models using Y and Z variables can either be over- or underestimated. Another consequence of the CIA is that statistically matched files are only suited for analyses of unconditional associations between Y and Z and associations conditional on only a subset of all possible confounding variables, that is, on only a subset of X variables. By design, Z and Y will be independent conditional on all X variables in the matched file.

Lacking additional information on $f(X, Y, Z)$, one approach to avoid the assumption of conditional independence is to perform sensitivity analyses (consider among others Kadane (1978); Moriarity and Scheuren (2001); D'Orazio et al. (2006a); D'Orazio et al.

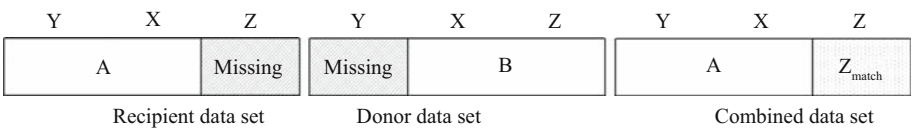

Fig. 2. Goal of micro approaches to statistical matching.

(2009); Conti et al. (2012, 2016); Rubin (1986); Rässler (2002, 2003); Rässler and Kiesl (2009)). These approaches typically utilize logical constraints to reduce uncertainty, for example, on the unknown correlation of Y and Z, ρ_{YZ}. In this case, the constraints follow from the necessity of (X, Y, Z)'s correlation matrix to be positive semidefinite and the fact that, apart from ρ_{YZ}, the correlation matrix can be estimated from A and B alone. Depending on the strength of the correlation between X and Y, and X and Z, the range of possible ρ_{YZ}s can be very small or not restricted at all.

An alternative approach for avoiding the assumption of conditional independence is to make use of additional available information. Singh et al. (1993), for example, provides a nonparametric micro approach (based on ideas in Paass (1985)) that can utilize auxiliary information in the form of a data set C in which X, Y and Z are jointly observed by first finding a nearest neighbor with respect to (X,Y) for each unit from A in C and donating their Z information to obtain (X, Y, Z_C). In a second step, Z_C is replaced by Z_B by finding a nearest neighbor with respect to (X, Z) in B. Other approaches include Bayesian methods, parametric, nonparametric and mixed approaches (for example, Kadane (1978); Paass (1985); Rässler (2003); Moriarity and Scheuren (2001, 2003); Filippello et al. (2004); Gilula et al. (2006); Gilula and McCulloch (2013); Fosdick et al. (2016)). Some utilize information on parameters regarding the distribution of Y and Z, others use C to estimate the conditional distribution of Y and Z given X. Again, we refer to D'Orazio et al. (2006b) for an overview.

3. Statistical Matching as a Supplement to Record Linkage

The goal of using statistical matching as a supplement to record linkage is to handle the missing data situation explained in Subsection 2.1 and depicted again using the statistical matching notation in the left-most panel of Figure 3. For consenting units, record linkage is performed in the usual way, while statistical matching is performed for all units that did not provide linkage consent. Note that in the statistical matching literature, A always denotes the data recipients, B denotes the donors, and C denotes the auxiliary data set in which all variables are jointly observed. Thus, to be consistent with this notation, A only comprises the survey data of the nonconsenters, B is still the donor data set, and C contains the combined data of the consenters in our context.

Besides conditional independence, there is another implicit assumption if traditional statistical matching is to be used as a supplement to record linkage. This assumption (described below) is necessary because the missing data situation generated by nonconsent to record linkage is different to the situation that statistical matching techniques are designed for. In the standard statistical matching scenario, the units in data set A and B are disjoint and X, Y, and Z are never jointly observed (D'Orazio et al. 2006b). The missingness of Z in A and of Y in B is therefore missing by design. If A and B are both

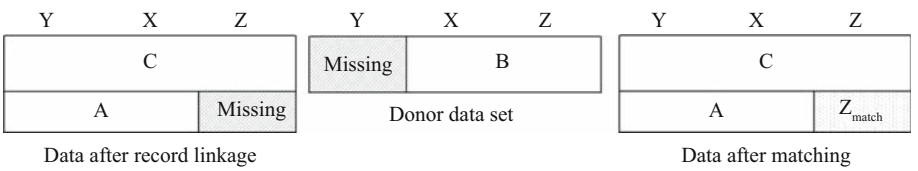

Fig. 3. Statistical matching as a supplement to record linkage

independent random samples of the population, the missing information in both files is missing completely at random (MCAR; Rubin (1976)).

The missingness generated by nonconsent to record linkage, on the other hand, is likely to not be MCAR. This implies that nonconsenters are not a random sample from the population. The partially observed realizations of the nonconsenters are random vectors of $f(x, y, z|\text{nonconsent})$ (with Z unobserved) which is not necessarily identical to $f(x, y, z)$. However, statistical matching does assume that both data sets A and B are random samples from the same distribution. If this is not true, it suffices to assume (at least for traditional statistical matching) that the conditional distribution of Z given X is the same in A and B (D'Orazio et al. 2006b). In our situation – assuming no selectivity in B – this translates to:

$$f(z|x, \text{nonconsent}) = f(z|x) \tag{2}$$

The distribution of Z conditional on X of the nonconsenters $f(z|x,\text{nonconsent})$ has to be the same as in B. This essentially means that one must assume the Z information for all nonconsenters is missing at random given X (MAR; Rubin (1976)). However, it is important to keep in mind that for Equation 2 to hold it is not sufficient that the probability to consent only depends on X. The selection mechanism for the complete process that leads to the final data of the consenters must only depend on X, that is, any selectivity introduced at the sampling stage, or because of nonconsent, must be fully explainable by X. This assumption can be seen as critical if only a small number of variables exist in X.

4. Problems in Practice

For statistical matching to be successful, the variables contained in X need to be measured similarly across the data sources (D'Orazio et al. 2006b; Meinfelder 2013). The main idea of statistical matching is to utilize the common variables X, and structural differences in the measurement of X in A, the recipients, and B, the donor data set, can therefore be highly problematic, for example, if the matching variables are measured with different levels of precision in the two data sources. Most importantly, the measurements should be free from bias, or, if bias exists, both sources need to be affected similarly. For example, with traditional statistical matching techniques, if X is biased differently in the recipient data set than in the donor data set, then the imputation will be based on the wrong value of X.

The assumptions regarding the bias behavior are especially problematic in the context of matching survey data with administrative records. Surveys are prone to measurement error since interviewers, question wording, memory of respondents, and various other factors can have effects on both accuracy and precision – bias and variance – of the measurement (Biemer et al. 2011). While administrative data can have different measurement problems (Oberski et al. 2017), the errors on the survey side alone can have detrimental effects on statistical matching even if the measurement in the administrative data is perfectly accurate and precise. For this reason, it is essential to identify all potential measurement differences in the two files and adjust the matching procedure accordingly.

Besides these very general remarks that apply to all statistical matching procedures, there are difficulties specific to only a subset of the available statistical matching techniques. Without going into extensive detail, we note that good parametric models are necessary to express the relationship between Z and X for all traditional statistical

matching methods, with the exception of nonparametric techniques like nearest neighbor. Similar to parametric imputation, parametric statistical matching is thus infeasible in the context of highly complex administrative data structures (see also our discussion at the end of Subsection 2.1).

Statistical matching techniques that utilize logical constraints are almost never used in practical statistical matching applications (Meinfelder 2013). The main reason is that they are only feasible if the number of variables within each of the vectors X, Y, and Z is relatively small. In addition, some methods make assumptions regarding the distribution of X, Y, and Z – the most prominent being multivariate normality. Given the complexity of variables X, Y and Z used in the context of merging survey and complex administrative data sets, with bounds, skip patterns, and logical constraints between the variables, such approaches are not feasible in applications similar to our setting. Besides, in many surveys most of the variables are discrete in nature or are measured on a discrete scale. Thus, the assumption of multivariate normality in particular, is often unrealistic. Furthermore, the uncertainty evaluation becomes much more complex if MCAR does not hold. The uncertainty is then a combination of the uncertainty of the missingness model and of the model parameter uncertainty (D'Orazio et al. 2006b).

In the supplement to record linkage scenario, there is auxiliary information available in the form of the successfully linked data of all consenting survey respondents. This could potentially be used as an auxiliary data set, C, for which X, Y and Z are jointly observed. Excluding parametric techniques for the same reason as above, to our knowledge the only nonparametric method proposed in the literature for incorporating C is the method by Singh et al. (1993) explained in Subsection 2.2. However, in settings like ours, it is very similar and offers essentially no benefit compared to nearest neighbor hot-deck imputation (cf. Subsection 2.1), which is essentially the first step of the method. When merging survey and administrative data, the donor pool is the complete population, which typically means that we will be able to find donors that match (almost) exactly on all the variables in X and Z_C. In this case, the second step of Singh et al. (1993) will not lead to any improvements, since Z_B will be equal to Z_C for all units. Therefore, the true donor pool will remain to be the records contained in C and the large pool in B cannot be utilized.

Given that methods that quantify the uncertainty from matching and methods that use auxiliary information cannot be exploited for our application for the reasons given above, we focus on traditional nearest neighbor techniques for the remainder of this article. Nearest neighbor methods are especially attractive in our case as they are nonparametric and thus are unaffected by the complexity of Z in administrative data sets.

We note that nearest neighbor hot-deck imputation (as explained in Subsection 2.1) has some similarities with statistical matching. The major difference is, data sets A and C are matched using both X and Y as matching variables instead of data sets A and B using only X. This means that with imputation we would not need to assume conditional independence. In addition, the missing at random assumption would be weakened to:

$$f(z|x, y, \text{nonconsent}) = f(z|x, y, \text{consent}) \tag{3}$$

However, as stated above, hot-deck methods are heavily dependent on the size of the donor pool and the donor-to-recipient ratio. While statistical matching can utilize the vast donor pool in the administrative data set B, imputation can only use the donors in C. This

means that if statistical matching can be used beneficially, it is more generally applicable than nearest neighbor hot-deck imputation, as it is independent of the consent rate and the sample size.

We conclude this section by noting that we do not believe that the conditional independence assumption and the missing-at-random assumption will ever be fully met in practice. However, we know that if we only use those cases that consented to the linkage of the data sources, we generally need to assume *consenting completely at random* if we want to get unbiased results. Arguably, this is also a rather strong assumption. Thus, the empirical question to answer is: would we be better off using only the data of the consenters, or could statistical matching be used to reduce the bias from assuming *consenting completely at random*? We do not expect to get completely unbiased results through statistical matching, but if the impacts of violating the statistical matching assumptions are minor, we might still be able to improve over the results based on using only the data of the consenters.

This reasoning is the motivation for the simulation studies described in the next sections.

5. Data Sources Used in the Evaluation Study

To evaluate whether statistical matching can be a viable supplement to record linkage, we use two large (and independent) panel surveys in Germany: the National Educational Panel Study and the Panel Study "Labour Market and Social Security". Both are linked to individual administrative process data from the German Federal Employment Agency. In our application, this administrative data set – the Integrated Employment Biographies – is used as the donor file B. The recipient file A consists of the nonconsenters of the National Educational Panel Study and the nonconsenters of the Panel Study "Labour Market and Social Security", respectively. We perform separate evaluation studies on both panel surveys. Before we discuss the design of these evaluation studies in more detail, this section provides a brief overview of the survey and administrative data sources.

5.1. Integrated Employment Biographies

The Integrated Employment Biographies (IEB) consists of administrative data obtained from social security notifications and different business processes of the German Federal Employment Agency. The different data sources are integrated for and by the Institute for Employment Research.

Figure 4 provides an overview of the business processes that generate IEB data. BeH information is provided for every employee covered by social security. Exclusions include individuals who did not enter the labor market and individuals who were self employed, since these groups are not subject to mandatory social security contributions. LeH and (X)LHG data are generated for individuals who received benefits in accordance with the Social Code Books (SGB) II (Sozialgesetzbuch 2003) and III (Sozialgesetzbuch 1997) (SGB II regulates welfare benefits for employable jobseekers in need and SGB III regulates employment promotion, in particular unemployment insurance). MTH and (X)ASU data are generated for individuals who were registered as jobseekers with the Federal Employment Agency or who participated in an employment or training program.

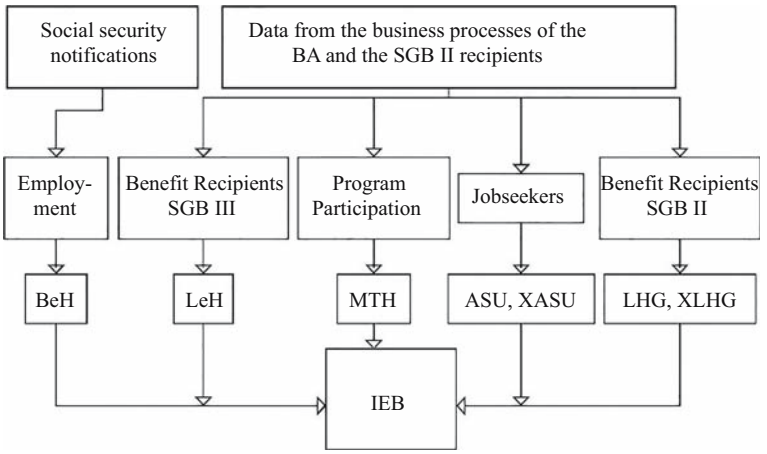

Fig. 4. Process data of the German Federal Employment Agency.

We refer to Jacobebbinghaus and Seth (2010) for a detailed description of the different data sources and of the IEB.

The IEB consists of a very large proportion of German residents, but not all. Thus, recipients of the statistical matching procedure should be limited to survey respondents who are also part of this subset of German residents. Note that in our evaluation study, this is guaranteed by design, as we only use the successfully linked cases.

Due to computational demands of the statistical matching procedure, it is mandatory to restrict the number of observations in B. Furthermore, researchers at the IAB cannot access the full IEB directly, since the size of the data set containing several billion records makes data handling difficult. For this reason the SIAB – a two-percent random sample from the IEB – is provided as a scientific use file that is easily accessible for all researchers at the IAB. Thus, we use the SIAB as the donor data set B for statistical matching. It still provides a very large donor pool of more than 1.7 million individuals and therefore guarantees a non-problematic donor to recipient ratio.

Availability and quality of IEB and SIAB data depend on various factors, including the data generating processes. It is out of scope of this article to go into further details (more information can be found in Antoni et al. (2016)). However, note that data on residents of federal states in the former German Democratic Republic are only available from 1993. Thus, to avoid biases, statistical matching is used on information generated after 1993. The information exclusive to the IEB, that is, the Z variables, mainly refer to individual employment history.

5.2. *National Educational Panel Study*

The National Educational Panel Study (NEPS) is carried out by the Leibniz Institute for Educational Trajectories at the University of Bamberg. The NEPS collects longitudinal data on competency development, educational processes, educational decisions and returns to education in Germany. Panel surveys on different age cohorts are conducted that provide data throughout the life course. The NEPS Starting Cohort 6 collects data on the

adult cohort. After a longer period between the first and second waves, which were carried out in 2007/2008 and 2009/2010, respectively, surveys for the adult cohort have been conducted yearly since 2011. The sample is drawn from municipality registration records of residents using a two-stage cluster sampling design with communities defining the primary sampling units and simple random sampling without replacement of individuals at the second stage. The target population of the adult cohort comprises residents in Germany who were born between 1944 and 1986, regardless of their nationality (Blossfeld et al. 2011). Variables that are exclusive to the NEPS data (that is, unavailable in the IEB/SIAB) are numerous. A unique characteristic of the NEPS compared to other surveys is the detail in information regarding the educational history of the respondents that form the Y variables of interest in the NEPS evaluation study.

Record linkage of NEPS and IEB data was carried out based on the nonunique identifiers, first and last name, date of birth, sex, and address information – postal code, city, street name, and house number. The consent rate in the NEPS adult cohort at the time of the linkage was 82% – yielding 14,065 consenters. Among the units that consented, 83.7%, that is, 11,778 units, could be linked deterministically and 7.5% (1,053 units) probabilistically. In the NEPS linkage, a link is called deterministic if the identifiers either match exactly or differ only in such a way that the probability for false positive links is still extremely low. For our evaluation study, we need a data set for which it is prudent to assume that all records are linked correctly. Therefore, we only keep those cases for which a deterministic linkage was possible. After additionally excluding every survey respondent whose latest linked IEB information is older than 1993, we arrive at a final data set consisting of 11,550 individuals. This data set is denoted as D^N_{det}.

5.3. *Panel Study "Labour Market and Social Security"*

The Panel Study "Labour Market and Social Security" (PASS) is an ongoing, nationally representative German household panel study, started in 2006 by the Institute for Employment Research. The aim of this study is to provide a database that enables an analysis of the dynamics of welfare benefits receipt after the introduction of the Unemployment Benefit II scheme in Germany in 2005. Information on labor market outcomes, household income, and unemployment benefit receipt are collected from more than 12,000 households annually. In addition to household interviews with the heads of the households, about 15,000 interviews with individual household members aged 15 and older are carried out.

The original PASS sample is composed of two subsamples: 1) a sample of households receiving unemployment benefit II (UB II Sample), which is drawn from recipient registers at the Federal Employment Agency; and 2) a sample of households from the general German population with an oversample of households with low economic status. The UB II Sample is refreshed each year to include new entries into the UB II population. PASS also introduced a replenishment sample for the general population sample in its fifth wave (for further information, see Trappmann et al. (2013)). As in the NEPS, there are many variables in the PASS that are not included in the administrative data of the IEB/SIAB. In particular, information on behaviors, attitudes, and subjective perceptions on the topics of social welfare benefits and labor market integration are available, which are the Y

variables of interest in the PASS evaluation study. The linkage consent rate to the IEB administrative data after the first five waves of PASS was at 79% (24,599 consenters). 87% (21,363 units) of the consenters could be successfully linked to the administrative data. 86% of the linkages were deterministic (18,425) using first and last name, date of birth, sex, and address information. Using our exclusion restriction that records need to be linked deterministically and that IEB spells need to be available after 1993, we end up with 18,202 individuals who are included in D^P_{det}.

6. Design of the Evaluation Study

We create synthetic nonconsent in the subset of deterministically linked respondents and check if, and to what extent, differences in estimates compared to before-deletion estimates can be reduced by using statistical matching as a supplement to record linkage. The design of our evaluation study comprises three steps. In the first step, we identify the deterministically linked cases in both data sources. In the second step, we model the probability of nonconsent based on the full survey data and use the predicted consent probabilities to introduce synthetic nonconsent among the true consenters. In the third step, we use statistical matching to find suitable administrative data donors for the generated nonconsenters and evaluate whether statistical matching reduces these differences. The three steps are visualized in Figure 5. Since statistical matching is only performed for the synthetic nonconsenters, they are denoted by A (the data recipients) in the figure, while the synthetic consenters are denoted by C, as record linkage is possible and thus all variables are available for them.

We note that although both surveys use complex sampling designs, we do not need to take any extra steps during the matching to account for the design, since both surveys are matched to a simple random sample of the IEB and the sampling design of the surveys is not relevant for nearest neighbor matching in this case. For statistical matching methods dealing with matching survey data sets with differing sampling designs, we refer the interested reader to Rubin (1986), Renssen (1998), Wu (2004) and Conti et al. (2016) for more details.

6.1. Generating Synthetic Nonconsent

In the data of all deterministically linked respondents D_{det}, the variable vector (X,Y,Z) is completely observed and the empirical distribution $f_{det}(x, y, z)$ of (X, Y, Z) is known (note that we always drop the superscripts N and P when we are not referring to a specific data source). These fully observed data will serve as the benchmark to evaluate whether nonconsent bias can be reduced by the proposed methodology. To introduce synthetic nonconsent among the true consenters based on realistic assumptions we use the full

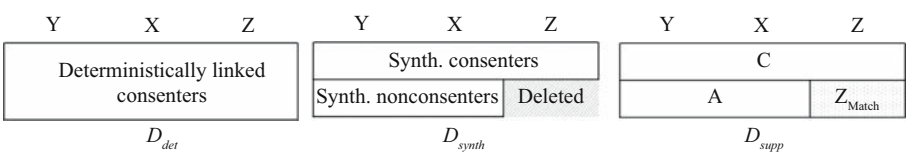

Fig. 5. The three steps of the evaluation study (left to right).

survey data to set up a model for the consent propensity. Specifically, we estimate a flexible, nonparametric logistic spline regression model with consent/nonconsent as the outcome variable and all X variables, as well as additional variables from Y for all survey respondents as covariates. All continuous covariates are included as B-splines. Since we need an estimated response propensity of every individual in D_{det}, survey variables used in the consent model that are subject to missingness need to be imputed. We use the software package mice in R (Van Buuren and Groothuis-Oudshoorn 2011) to generate m=5 imputations (based on eight iterations) using predictive mean matching and classification trees for metric and categorical variables, respectively. Hence, we need to make the implicit assumption that the missingness mechanism for all the imputed variables is missing at random.

Following the multiple imputation framework, the predicted consent propensities are obtained by averaging the predictions of the consent model from each of the imputed data sets. Respondents in D_{det} are stochastically chosen to be synthetic nonconsenters with a probability equal to their estimated consent propensity. Their Z variables are deleted to form D_{synth}.

In the case of the NEPS and the PASS, the resulting synthetic nonconsent rates are roughly 15 and 13%, respectively. To evaluate the effects of nonconsent and the performance of statistical matching under various assumptions, we also created data sets with synthetic nonconsent rates of 40 and 60% by adjusting the nonconsent probabilities accordingly. The observed empirical distribution of (X,Y,Z) for the remaining consenters is denoted as $f_{synth}(x, y, z)$.

Note that we implicitly make the assumption that the consent mechanism is *consenting at random* with respect to X and Y, that is, the probability to consent only depends on the survey variables. This assumption is necessary in our evaluation setup, since the true Z values are not observed for the nonconsenters by definition. The nonconsent model can therefore only include survey variables. Thus, the findings from our study will only be generalizable to situations where this assumption regarding the consent mechanism holds. However, it is prudent to assume that if statistical matching performs poorly in our evaluation study, this will also be the case if the assignment mechanism also depends on Z.

Since the data of the synthetic consenters are a random subsample of D_{det}, we cannot determine directly whether differences between the estimate using the synthetic consenters only and the estimate based on D_{det} are systematic or due to chance. A first indicator of potential bias in the estimates using only the synthetic consenters would be if any of the coefficients in the consent propensity model are significant. If there are significant parameters in the model, which is true in our case, then the assumption that the consent mechanism is consenting completely at random (CCAR) likely does not hold. As an additional evaluation whether, and to what degree, the observed differences in our study are systematic, we provide 95% confidence intervals for the estimates of interest under the null hypothesis that the consent mechanism is CCAR. Thus, the confidence intervals computed under the null hypothesis will be a measure of how much additional uncertainty we might expect due to the reduced sample size because of synthetic nonconsent. We can use these confidence intervals for classical hypothesis testing. If the confidence interval does not include the estimate of interest obtained using the remaining consenters based on the model described above, the null hypothesis that the consenting process is CCAR can be rejected.

The confidence intervals are obtained using Monte Carlo simulations. To obtain the confidence interval for an estimate of interest given a specific synthetic nonconsent rate r, we randomly delete $r \times 100\%$ of the data in D_{det} and compute the estimate of interest, based on the remaining cases. This is a realization of the estimand under the null hypothesis. By repeating this process 5,000 times, we make certain that the resulting empirical distribution is a good approximation of the true distribution under the null hypothesis. 95% confidence intervals are obtained by searching for the 2.5% and 97.5% quantiles of this distribution.

Note that we use Monte Carlo simulations only to create the confidence intervals – D_{synth} is created only once. This is a limitation of this evaluation study, since the creation of D_{synth} is subject to randomness and thus the results could differ over repeated simulation runs. However, due to the numerical expensiveness of the matching procedure, we are limited to a single run for the actual matching.

6.2. *Statistical Matching for all Synthetic Nonconsenters*

With the aim of reducing the nonconsent bias, statistical matching is performed for all synthetic nonconsenters. The specific matching method used here is called random distance hot-deck matching (D'Orazio et al. 2006b). For every synthetic nonconsenter, the method finds those k individuals from the administrative database who have the lowest distance regarding X, and from these k records, selects one at random and uses it as a donor. The main idea is that the empirical distribution of the k nearest neighbors' Z values approximates the posterior predictive distribution of missing data in Z given the survey respondent's realized X value, and the approach takes a random sample of size one from this conditional distribution. Similarly to stochastic versus deterministic imputation, it is preferable to draw from the posterior predictive distribution instead of simply using the expected value (Little and Rubin 2002). One could pick more than one donor in the spirit of multiple imputation to fully reflect the uncertainty that comes from matching randomly among the k closest donors (Rubin 1978, 1987). However, this approach is computationally intensive and is unlikely to affect the differences in point estimates since parameter estimates after multiple imputation are just averages over the parameter estimates in all imputed data sets. Nonetheless, as a sensitivity check we evaluated whether our findings change if we used $m = 5$ donors for each record. Since we did not find any differences in the results, the results reported below are based on picking only one donor.

We use the standardized Euclidean distance as a distance measure and set k to 20. While the Mahalanobis distance should do a better job for most statistical matching purposes, it is computationally more expensive. In addition, due to the large donor pool contained in the SIAB, the benefits of the Mahalanobis compared to the Euclidean distance should be negligible, since matching will be almost exact for most survey respondents.

The following variables are used as matching variables X: an indicator of whether the individual was ever married, age, an indicator for having children, salary in 2010, occupation, place of residence (formerly West or East Germany), and three variables on whether BeH, LeH, and LHG information is available. Some of these matching variables are only available for specific individuals due to the different data-generating processes in the IEB. Sex and nationality (German yes/no) are used as blocking variables, that is, IEB

units are excluded as potential donors for survey respondents if they do not have identical values in these variables. All of these variables are used in the matching procedure of both PASS and NEPS data to allow a comparison of the results in the two case studies. More information on the variables used can be found in Section 9, Appendix.

After statistical matching, we add the Z information of the identified matches for all synthetic nonconsenters to D_{synth}, and thus obtain a data set D_{supp}, for which (X, Y, Z) is again available for all deterministically linked survey respondents. The resulting empirical distribution is denoted as $f_{supp}(x, y, z)$. We can then evaluate whether differences in $f_{synth}(x, y, z)$ compared to $f_{det}(x, y, z)$ are reduced in $f_{supp}(x, y, z)$. Specifically, we look at marginal distributions in Z variables, correlations between Y and Z variables, and coefficients of regression models that use both Y and Z variables. For ease of reading, we use the terms *reference* or *benchmark* estimate for the estimates of interest using D_{det}.

7. Results

Using the predicted consent probabilities directly induces almost no differences in estimates in D_{synth} compared to the reference. After increasing the nonconsent rate to 40%, large differences can be observed for some estimands. Increasing the nonconsent rate to 60% increases these differences further. However, the general findings regarding the bias and the success of the statistical matching approach are similar for both consent rates. Therefore, we will only present the results using the smaller and more realistic nonconsent rate of 40% in this section.

7.1. Marginal Distributions and Means

In principle, marginal distributions for administrative data variables are available for the population in the complete administrative data. This means that data linkage would not be necessary to begin with. Thus, one could argue that biases in these marginal distributions should not be of any concern. However, this argument is only valid if the population of interest and the population of the administrative data are actually the same. If the survey population is a subset of the population of the administrative data, as for example in the case of the PASS subsample of unemployment benefit II recipients, it will still be important to evaluate whether the proposed method helps to correct for nonconsent bias in marginal distributions. However, note that the assumption that the conditional distribution $f(z|x)$ is the same for units in A and B is stronger if the survey population and administrative data population are different.

In our evaluation study, we examine the marginal distributions of some key measures in the IEB. An important characteristic of the IEB is its accurate information regarding the employment history of each individual (Jacobebbinghaus and Seth 2010). Thus, we evaluate whether statistical matching can reduce nonconsent bias in marginal distributions of these Z variables. Figure 6 presents results regarding the means of three important Z variables from the IEB: time in employment, complete gross salary earned in 2011, and the complete duration of Unemployment Benefit I receipt. Unemployment Benefit I is a specific social welfare payment in Germany that is paid during the first 6 to 18 months of unemployment. All values depicted here are ratios of the respective means to their benchmark values, that is, to the means in the complete linked data set D_{det}^{N}. As a reference,

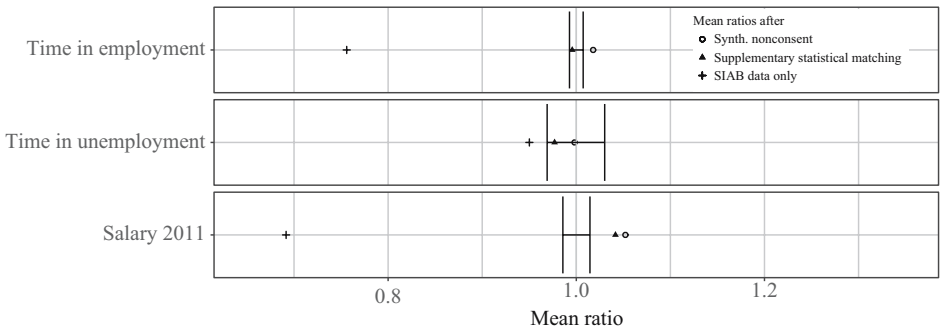

Fig. 6. Estimated means of three IEB variables divided by their NEPS benchmark estimate. The ratios are computed 1) after inducing 40% synthetic nonconsent and 2) after subsequent matching (ratios for the SIAB are included as a reference). The bars indicate the bounds of the 95% confidence intervals for the respective mean under the assumption of consenting completely at random.

Figure 6 also contains this ratio for the variables in B (the SIAB). Furthermore, a 95% confidence interval for the estimates assuming *consenting completely at random* is provided.

In the case of the NEPS, the synthetic nonconsent generates systematic differences in the variables time in employment and salary in 2011. In both cases, the confidence interval for the respective mean under the assumption of *consenting completely at random* does not cover the mean after inducing synthetic nonconsent. The difference in both variables can be reduced by using subsequent statistical matching for all synthetic nonconsenters. In contrast, no differences are created by the synthetic nonconsent process for the total duration of Unemployment Benefit I receipt and subsequent statistical matching slightly worsens the estimate from a bias perspective. The estimates based on the matched data are always close to the benchmark value despite the fact that estimates using the SIAB data would be substantially different, especially for time in employment and salary in 2011.

Looking at the mean ratios for the PASS (Figure 7), the findings are similar for the employment-related variables. For the variable salary 2011, the difference is slightly larger after supplementing the data with statistical matches.

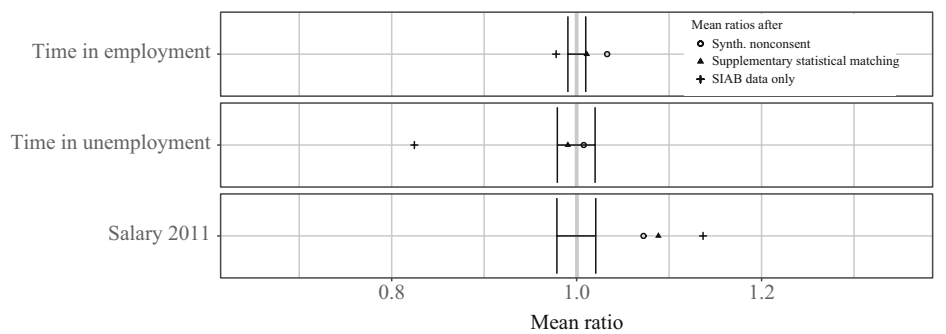

Fig. 7. Estimated means of three IEB variables divided by their PASS benchmark estimate. The ratios are computed 1) after inducing 40% synthetic nonconsent and 2) after subsequent matching (ratios for the SIAB are included as a reference). The bars indicate the bounds of the 95% confidence intervals for the respective mean under the assumption of consenting completely at random.

7.2. Correlations and Regression Model Parameters

The goal for record linkage is to be able to analyze the combined data. Univariate analyses can be performed on the administrative data without linkage (though, not always with respect to the specific survey population). Thus, it is essential to evaluate the methodology for bivariate and multivariate estimands that utilize both Y and Z variables. We only present results for the NEPS data in this section. Results obtained from the PASS data showed similar patterns and thus we exclude them for brevity.

In Figure 8, we present the effects of statistical matching on correlations of the three aggregate administrative data variables introduced in the previous section with three Y variables, that is, variables that are only available in the survey: years of schooling, age at first employment, and length of first employment. The before-deletion correlation – the empirical correlation in D^N_{det} – is plotted against the observed correlations in the data sets after creating synthetic nonconsent and using statistical matching for all synthetic nonconsenters.

We observe that there are more or less no differences in estimates using D_{synth} compared to the benchmark. However, almost all estimates are shrunk towards zero if statistical matching is applied. Instead of correcting for any differences created by nonconsent, statistical matching actually increases these differences. As explained in Subsection 2.2, the shrinkage towards zero is an indication that the conditional independence assumption is invalid.

We also estimate two regression models to evaluate to what extent multivariate relationships can be preserved after statistical matching; first, a Cox proportional hazards model (Cox and Oakes 1984) of the (log) length of the first unemployment episode as the dependent variable, and second, a linear regression model of the (log) gross salary in 2011

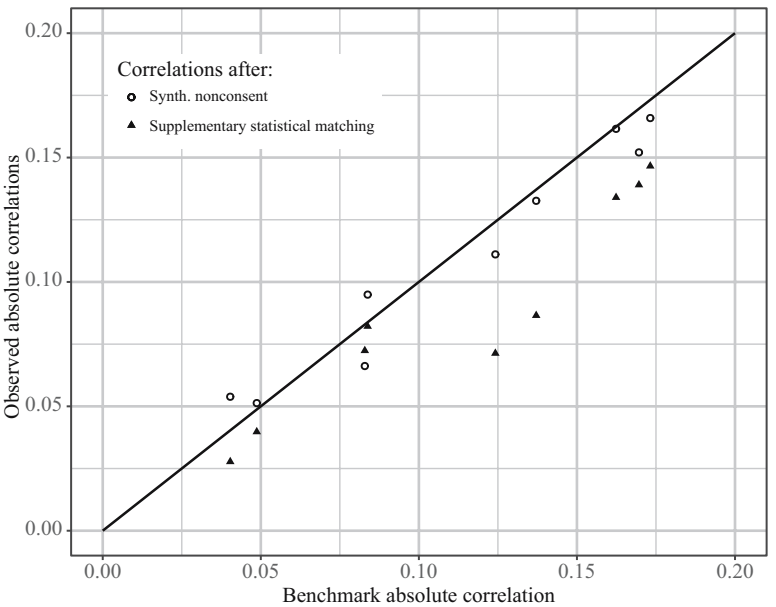

Fig. 8. Bivariate correlations for a subset of survey and administrative variables after 40% synthetic nonconsent and subsequent matching based on the NEPS data compared to the benchmark correlations.

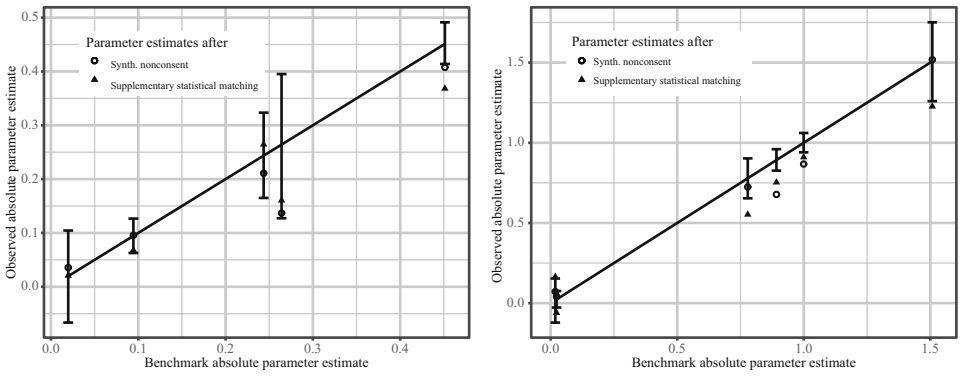

Fig. 9. Regression parameter estimates after 40% synthetic nonconsent and subsequent matching based on the NEPS data compared to the benchmark estimates (the bars indicate the bounds of the 95% confidence intervals for the respective parameter estimate under the assumption of consenting completely at random).

as the dependent variable. Both models use sex, age (and a quadratic term of age in the case of the salary model), occupational training, number of school years, and whether the mother's mother tongue is German as independent variables.

Similar to Figure 8, Figure 9 plots the before-deletion parameter estimates against the parameter estimates after inducing synthetic nonconsent and subsequent supplementation with statistical matching. The results of the parameter estimates are less cohesive. There are only a couple of coefficients for which differences in estimates after synthetic nonconsent compared to the reference are substantial. Again, the confidence intervals provide a test of whether or not the synthetic nonconsent process is significantly different from consenting completely at random. Point estimates are in some cases closer to the reference value after statistical matching, but in other cases the differences are larger. Also, absolute values of parameter estimates after supplementation are sometimes lower and sometimes higher than the true values. The unconditional relationship between Y and Z variables that can be observed in the matched data set is only the part of the relationship that can be explained by X. As explained in Subsection 2.2, omitting important confounding variables in X can lead to overestimation, as well as underestimation, of the unconditional effect after statistical matching. Also, even if all important confounding variables are included as matching variables, the lack of a potentially existing effect – conditional on every confounding variable – in the matched data can also lead to underestimation of the absolute value of the regression coefficients after matching. Both of these problems are only relevant if the conditional independence assumption is violated, but as our results concerning the correlations suggest, this is the case for almost all pairs of Y and Z variables that we examined.

8. Conclusion

Supplementing record linkage of survey and administrative data with nearest neighbor statistical matching is a straightforward idea when trying to reduce nonconsent biases. Since good parametric models are not necessary for nearest neighbor techniques and donor

sparseness is never an issue if large administrative data sets are used as the donor pool, it is the most widely applicable method available. However, the assumptions that are implicit when traditional statistical matching is used as a supplement to record linkage are very strong and will most likely never hold completely. The goal of the simulation study presented in this article was to evaluate empirically how well nearest neighbor matching performs, despite these assumptions. Our results suggest that biases in marginal distributions of administrative data variables can be corrected quite well, depending on the predictive power of the matching variables for the variable of interest. This is a particularly useful finding for situations where marginal distributions of administrative variables are desired for the survey population under study. However, the method is less suited for more complex analyses. The implications of the violation of the conditional independence assumption were substantial for both bivariate and multivariate analyses on the supplemented data sets.

Another downside of the approach is that the seemingly simple matching problem turns into a tedious task in practice, since preparing multiple data sources for statistical matching is a time-consuming and resource intensive process. In this study, significant efforts were undertaken to implement a high quality statistical matching procedure and analysis. Multiple issues, most of them related to measurement differences in the two data sets had to be dealt with in advance. These differences are likely to be present in any application of statistical matching of survey and administrative data.

The results from our simulations suggest that – with the exception of marginal distributions – the problems created by statistically matching nonconsenting units are worse than ignoring the nonconsent problem. Thus, even though the results are not easily generalizable to other applications, we advise caution when using nearest neighbor statistical matching to reduce linkage nonconsent bias for more complex estimates.

If other statistical agencies are considering statistical matching as a supplement to record linkage, our simulation design can be seen as a roadmap to empirically evaluate whether biases from nonconsent can be reduced for the specific application at hand. We emphasize that it is generally impossible to derive analytically which assumptions are stronger: the consenting completely at random assumption implied when analyzing only the data of the consenters or the assumptions required for statistical matching as discussed in Subsection 2.2 and Section 3. Both assumptions will never be fully met in practice, but the impact of the violation of the assumptions will depend on the available data, the nonconsent process, and the analysis of interest. Thus, statistical agencies might follow the simulation setup laid out in Section 6 to decide whether statistical matching could be a useful tool for their analysis goals, especially if a rich pool of jointly observed variables is available. We note that our evaluation study focused only on biases. Further research could extend our approach by including appropriate procedures based on the multiple imputation framework for enabling valid variance estimates after matching.

As discussed in Section 4, the missing at random assumption necessary for nearest neighbor imputation is weaker than for supplemental statistical matching. Therefore, one area of future research could focus on nearest neighbor imputation as a method to reduce nonconsent bias in a similar evaluation setting. Additional research questions related to how analyses on the imputed data set will be influenced by poor donor to recipient ratios due to low consent rates, should also be explored in this context.

Table 1. Variables used in the Matching Procedure of NEPS/PASS and IEB Data.

Variable	Type	Prereq.	NEPS Mean (SD)	PASS Mean (SD)	IEB Mean (SD)
Sex (female = 1)	Blocking	-	.504	.527	.480
Nationality (ger = 1)	Blocking	-	.955	.906	.868
BEH Spell	Asymm. Block.	-	.915	.579	.889
LEH Spell	Asymm. Block.	-	.208	.287	.404
No LEH Spell	Asymm. Block.	-	.431	.281	.596
Fulltime Employed	Asymm. Block.	BEH	.855	.489	.761
Single (no = 1)	Matching	LEH	.602	.444	.452
Has Children	Matching	LEH	.747	.536	.385
Place of residence	Matching	-	.156	.229	.156
Year of birth	Matching	-	1,963.7 (11.3)	1,965.6 (16.5)	1,965.2 (17)
Income (adjusted)	Matching	BEH	2,630 (3,248)	1,839 (1,682)	2,009 (2,568)
Occupation 1	Matching	BEH	.020	.030	.025
Occupation 2	Matching	BEH	.0004	.0006	.0008
Occupation 3	Matching	BEH	.170	.281	.229
Occupation 4	Matching	BEH	.072	.041	.053
Occupation 5	Matching	BEH	.731	.627	.686
Occupation 6	Matching	BEH	.006	.020	.006

Means and standard deviations for continuous variables; proportions for dichotomous matching variables.
Income values are inflation adjusted to the level of 2010.
Occupation is a nominally scaled variable with six categories.

9. Appendix

Table 1 shows all variables that are used in the matching procedure. We categorize every variable into blocking, asymmetric blocking, and matching variables. If a variable is used as a blocking variable, only individuals in the administrative data who have identical values in this variable are allowed to be used as donors. Matching variables are used to compute the Euclidean distance. Asymmetric blocking variables are used if blocking is only possible for specific respondents.

To illustrate, in our application, asymmetric blocking is required for the following reason: the IEB combines different sources of data that are generated from different BA business processes (see Figure 4) and all data sources provide different information. The variables from the different sources can only be used to find a donor for a survey respondent if it is certain that this information should be available in this respondent's (and therefore every similar individual's) administrative data. Therefore, we have to find proof – or at least strong indicators – in the survey data that this BA process should have been initiated by the respondent. However, not finding these indicators does not necessarily mean that the respondent's administrative data does not include this information. Therefore, these indicators are used in the matching process as asymmetric blocking variables.

10. References

Andridge, R.R. and R.J. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78(1): 40–64.

Antoni, M., A. Ganzer, and P. vom Berge. 2016. *Sample of Integrated Labour Market Biographies (SIAB) 1975–2014*. FDZ-Datenreport 4, Institute for Employment Research, Nuremberg, Germany. Avaiable at: http://doku.iab.de/fdz/reporte/2016/DR_04-16_EN.pdf.

Antoni, M. and S. Seth. 2011. *ALWA-ADIAB – linked individual survey and administrative data for substantive and methodological research*. FDZ-Methodenreport 12, Institute for Employment Research, Nuremberg, Germany. Avaiable at: http://doku.iab.de/fdz/reporte/2011/DR_05-11.pdf.

Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman. 2011. *Measurement Errors in Surveys*. John Wiley & Sons.

Blossfeld, H.-P., H-G. Roßbach, and J. Von Maurice. 2011. "Education as a Lifelong Process." *Zeitschrift für Erziehungswissenschaft Sonderheft* 14. ISBN: 978-3-531-17785-4.

Brick, J.M. and G. Kalton. 1996. "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5(3): 215–238. Doi: http://dx.doi.org/10.1177/096228029600500302.

Brücker, H., M. Kroh, S. Bartsch, J. Goebel, S. Kühne, E. Liebau, P. Trübswetter, I. Tucci and J. Schupp. 2014. "The New IAB-SOEP Migration Sample: An Introduction into the Methodology and the Contents." *SOEP Survey Papers 216*. Avaiable at: http://hdl.handle.net/10419/103964.

Calderwood, L. and C. Lessof. 2009. "Enhancing Longitudinal Surveys By Linking to Administrative Data." In *Methodology of Longitudinal Surveys*, edited by P. Lynn, 55–72. New York: Wiley. ISBN: 978-0-470-01871-2.

Chen, J. and J. Shao. 2000. "Nearest Neighbor Imputation for Survey Data." *Journal of Official Statistics* 16(2): 113–131. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nearest-neighbor-imputation-for-survey-data.pdf.

Conti, P.L., D. Marella and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching." *Journal of Official Statistics* 28(1): 69–88. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/uncertainty-analysis-in-statistical-matching.pdf.

Conti, P.L., D. Marella and M. Scanu. 2016. "Statistical Matching Analysis for Complex Survey Data with Applications." *Journal of the American Statistical Association* 111(516): 1715–1725. Doi: http://dx.doi.org/01621459.2015.1112803.

Cox, D.R. and D. Oakes. 1984. *Analysis of Survival Data*. CRC Press.

da Silva, M.E.M., C.M. Coeli, M. Ventura, M. Palacios, M.M.F. Magnanini, T.M.C.R. Camargo and K.R. Camargo. 2012. "Informed Consent for Record Linkage: A Systematic Review." *Journal of Medical Ethics* 38(10): 639–642. Doi: http://dx.doi.org/10.1136/medethics-2011-100208.

D'Orazio, M., M. Di Zio and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty using Logical Constraints." *Journal of Official Statistics* 28(1): 137–157. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-for-categorical-data-displaying-uncertainty-and-using-logical-constraints.pdf.

D'Orazio, M., M. Di Zio and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. John Wiley & Sons.

D'Orazio, M., M. Di Zio and M. Scanu. 2009. "Uncertainty Intervals for Nonidentifiable Parameters in Statistical Matching." Proceedings of the 57th session of the International Statistical Institute, August 16–22, 2009, Durban, South Africa.

Fellegi, I.P. and A.B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64(328): 1183–1210. Doi: http://dx.doi.org/10.1080/01621459.1969.10501049.

Filippello, R., U. Guarnera and G. Jonas Lasinio. 2004. "Use of auxiliary information in statistical matching." Proceedings of the XLII Conference of the Italian Statistical 9–11 June 2014, Bari, Italy: 37–40.

Fosdick, B.K., M. DeYoreo and J.P. Reiter. 2016. "Categorical Data Fusion using Auxiliary Information." *The Annals of Applied Statistics* 10(4): 1907–1929. Doi: http://dx.doi.org/10.1214/16-AOAS925.

Fulton, J.A. 2012. *Respondent Consent to Use Administrative Data*, Ph. D. thesis, University of Maryland.

GDPR. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." *Official Journal of the European Union* L119: 1–88. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Gilula, Z. and R. McCulloch. 2013. "Multi Level Categorical Data Fusion using Partially Fused Data." *Quantitative Marketing and Economics* 11(3): 353–377. Doi: http://dx.doi.org/10.1007/s11129-013-9136-0.

Gilula, Z., R.E. McCulloch and P.E. Rossi. 2006. "A Direct Approach to Data Fusion." *Journal of Marketing Research* 43(1): 73–83. Doi: http://dx.doi.org/10.1509/jmkr.43.1.73.

Herzog, T.N., F.J. Scheuren and W.E. Winkler. 2007. *Data Quality and Record Linkage Techniques*. Springer Science & Business Media.

Jacobebbinghaus, P. and S. Seth. 2010. *Linked-Employer-Employee-Daten des IAB: LIAB – Querschnittmodell 2, 1993–2008*. FDZ-Datenreport, Institute for Employment Research, Nuremberg, Germany.

Jenkins, S.P., L. Cappellari, P. Lynn, A. Jäckle and E. Sala. 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. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00417.x.

Kadane, J.B. 1978. "Some Statistical Problems in Merging Data Files." *Compendium of Tax Research*, 159–179, Reprint in Journal of Official Statistics 17(3): 423–433. Avaiable at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/some-statistical-problems-in-merging-data-files.pdf.

Kreuter, F., J.W. Sakshaug and R. Tourangeau. 2016. "The Framing of the Record Linkage Consent Question." *International Journal of Public Opinion Research* 28(1): 142–152. Doi: http://dx.doi.org/10.1093/ijpor/edv006.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*, (2nd ed.). John Wiley & Sons.

Meinfelder, F. 2013. "Datenfusion: Theoretische Implikationen und praktische Umsetzung." In *Weiterentwicklung der amtlichen Haushaltsstatistiken*, edited by T. Riede, N. Ott and S. Bechthold, 83–98. Berlin: GWI Wissenschaftspolitik Infrastrukturentwicklung.

Moriarity, C. and F. Scheuren. 2001. "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure." *Journal of Official Statistics* 17(3): 407–422. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-a-paradigm-for-assessing-the-uncertainty-in-the-procedure.pdf.

Moriarity, C. and F. Scheuren. 2003. "A Note On Rubin's Statistical Matching using File Concatenation." *Journal of Business and Economic Statistics* (21): 65–73. Doi: http://dx.doi.org/10.1198/073500102288618766.

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. Doi: http://dx.doi.org/10.1080/13645579.2015.1019264.

Ness, A.R. 2004. "The Avon Longitudinal Study of Parents and Children (ALSPAC) – A Resource for the Study of the Environmental Determinants of Childhood Obesity." *European Journal of Endocrinology* 151(Suppl 3): U141–U149. Doi: http://dx.doi.org/10.1530/eje.0.151u141.

Oberski, D.L., A. Kirchner, S. Eckman and F. Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models."

Journal of the American Statistical Association. Doi: http://dx.doi.org/10.1080/01621459.2017.1302338.

Paass, G. 1985. "Statistical Record Linkage Methodology: State of the Art and Future Prospects." *Bulletin of the International Statistical Society. Proceedings of the 45th Session*. Voorburg, Netherlands: ISI.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer Science & Business Media.

Rässler, S. 2003. "A Non-Iterative Bayesian Approach to Statistical Matching." *Statistica Neerlandica* 57(1): 58–74. Doi: http://dx.doi.org/10.1111/1467-9574.00221.

Rässler, S. and H. Kiesl. 2009. "How Useful are Uncertainty Bounds? Some Recent Theory with an Application to Rubin's Causal Model." Proceedings of the 57th Session of the International Statistical Institute, August 16–22, 2009, Durban, South Africa. Available at https://www.isi-web.org/index.php/publications/proceedings.

Renssen, R.H. 1998. "Use of Statistical Matching Techniques in Calibration Estimation." *Survey Methodology* 24: 171–184. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/1998002/article/4354-eng.pdf.

Rodgers, W.L. 1984. "An Evaluation of Statistical Matching." *Journal of Business & Economic Statistics* 2(1): 91–102. Doi: http://dx.doi.org/10.1080/07350015.1984.10509373.

Rubin, D.B. 1976. "Inference and Missing Data." *Biometrika* (3): 581–592. Doi: http://dx.doi.org/10.2307/2335739.

Rubin, D.B. 1978. "Multiple Imputation in Sample Surveys – a Phenomological Bayesian Approach to Nonresponse." *Proceedings of the Survey Research Method Section of the American Statistical Association*: Joint Statistical Meetings 1978, San Diego, U.S.A.: 20–30. Available at: http://www.asasrms.org/Proceedings/index.html.

Rubin, D.B. 1986. "Statistical Matching using File Concatenation with Adjusted Weights and Multiple Imputations." *Journal of Business & Economic Statistics* 4(1): 87–94. Doi: http://dx.doi.org/10.1080/07350015.1986.10509497.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Sakshaug, J.W., M.P. Couper, M.B. Ofstedal and D.R. Weir. 2012. "Linking Survey and Administrative Records: Mechanisms of Consent." *Sociological Methods & Research* 41(4): 535–569. Doi: http://dx.doi.org/10.1177/0049124112460381.

Sakshaug, J.W. and M. Huber. 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. Doi: http://dx.doi.org/10.1093/jssam/smv034.

Sakshaug, J.W., S. Hülle, A. Schmucker and S. Liebig. 2017. "Exploring the Effects of Interviewer- and Self-administered Survey Modes on Record Linkage Consent Rates and Bias." *Survey Research Methods* 11(forthcoming): 171–188. Doi: http://dx.doi.org/10.18148/srm/2017.v11i2.7158.

Sakshaug, J.W. and F. Kreuter. 2012. "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data." *Survey Research Methods* 6(2): 113–122. Doi: http://dx.doi.org/10.18148/srm/2012.v6i2.5094.

Sakshaug, J.W. and B. Vicari. 2017. "Obtaining Record Linkage Consent from Establishments: The Impact of Question Placement on Consent Rates and Bias." *Journal of Survey Statistics and Methodology*. Doi: http://dx.doi.org/10.1093/jssam/smx009.

Sala, E., J. Burton and G. Knies. 2012. "Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics." *Sociological Methods & Research* 41(3): 414–439. Doi: http://dx.doi.org/10.1177/0049124112457330.

Schulte Nordholt, E., J. Van Zeijl and L. Hoeksma. 2014. *Dutch Census 2011, Analysis and Methodology*, Technical report, Statistics Netherlands. ISBN: 978-90-357-1948-4. Available at: https://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/0/2014b57pub.pdf.

Sims, C. 1972. "Comments on Okner (1972)." *Annals of Economic and Social Measurement* (1): 343–345.

Singh, A., H. Mantel, M. Kinack and G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption." *Survey Methodology* 19(1): 59–79. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199300114475.

Sozialgesetzbuch. 1997. SGB Drittes Buch (III) – "Arbeitsförderung".

Sozialgesetzbuch. 2003. SGB Zweites Buch (II) – "Grundsicherung für Arbeitsuchende".

Trappmann, M., J. Beste, A. Bethmann and G. Müller. 2013. "The PASS Panel Survey After Six Waves." *Journal for Labour Market Research* 46(4): 275–281. Doi: http://dx.doi.org/10.1007/s12651-013-0150-1.

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "MICE: Multivariate Imputation By Chained Equations in R." *Journal of Statistical Software* 45(3). Doi: http://dx.doi.org/10.18637/jss.v045.i03.

Wu, C. 2004. "Combining Information from Multiple Surveys through the Empirical Likelihood Method." *Canadian Journal of Statistics* 32(1): 15–26. Doi: http://dx.doi.org/10.2307/3315996.

Received June 2017
Revised May 2018
Accepted June 2018