Data Observer

Holger Lüthen, Carsten Schröder*, Markus M. Grabka, Jan Goebel, Tatjana Mika, Daniel Brüggmann, Sebastian Ellert and Hannah Penz

SOEP-RV: Linking German Socio-Economic Panel Data to Pension Records

https://doi.org/10.1515/jbnst-2021-0020
Received June 1, 2021; accepted June 14, 2021

Abstract: The aim of the project SOEP-RV is to link data from participants in the German Socio-Economic Panel (SOEP) survey to their individual Deutsche Rentenversicherung (German Pension Insurance) records. For all SOEP respondents who give explicit consent to record linkage, SOEP-RV creates a linked dataset that combines the comprehensive multi-topic SOEP data with detailed cross-sectional and longitudinal data on social security pension records covering the individual's entire insurance history. This article provides an overview of the record linkage project, highlights potentials for analysis of the linked data, compares key SOEP and pension insurance variables, and suggests a re-weighting procedure that corrects for selectivity. It concludes with details on the process of obtaining the data for scientific use.

Keywords: consent, pension records, record linkage, SOEP, SOEP-RV

JEL Classification: C89, C18

1 Introduction

Record linkage is a method for precisely matching microdata from different sources with the goal of expanding the potential of the data for research (e.g., Schnell 2014). Data linkage offers several benefits: In addition to broadening the

***Corresponding author: Carsten Schröder**, SOEP at DIW Berlin and Freie Universitaet Berlin, Berlin, Germany, E-mail: CSchroeder@diw.de
Holger Lüthen, Federal Ministry for Economic Affairs and Energy, Berlin, Germany
Markus M. Grabka, Jan Goebel and Hannah Penz, SOEP at DIW Berlin, Berlin, Germany
Tatjana Mika, Daniel Brüggmann and Sebastian Ellert, Deutsche Rentenversicherung Bund, Research Data Centre of German Pension Insurance, Berlin, Germany

range of variables and the observed temporal horizon, it provides opportunities for cross-validation of information and reduces the time burden on respondents. This paper describes the data linkage project SOEP-RV, which is being conducted by the German Socio-Economic Panel (SOEP) in partnership with the Research Data Centre of the German Pension Insurance (FDZ-RV). The aim of SOEP-RV is to link the SOEP data to administrative pension records. We do so by obtaining the social security numbers of consenting SOEP respondents from the statutory pension insurance, and using these—in adherence to the highest data security standards—for one-to-one linkage of the SOEP and pension insurance data.

The SOEP survey, established in 1984, is a multi-topic household panel study providing individual- and household-level information (see Goebel et al. 2019). Within the project, the SOEP is linked with two administrative datasets: (a) RTBN (Rentenbestand), a cross-sectional dataset that provides information on retirees' pension accounts pension stocks; and (b) VSKT (Versicherungskontenstichprobe), a longitudinal dataset in spell form that is comprised of an individual's insurance history from the age of 14–67. SOEP's main advantage is the broad set of variables it provides for the resident population of Germany, both at the individual and household level, including individual relationships within and between surveyed households. The administrative data add comprehensive social security information virtually without measurement error on a monthly level.

SOEP-RV expands the research potential of SOEP data in several respects: First, it broadens the range of variables available for analysis. For a number of specific pension types, the administrative data clearly exceed the SOEP's level of detail. Second, SOEP-RV extends the SOEP's biographical information beyond the time of the initial survey and provides supplementary information that can be used to fill in gaps that occurred due to nonresponse (Frick and Grabka 2005) or to correct for recall bias (Bound et al. 2001). The pension records add biographical social security information starting earliest at the age of 14 for all SOEP respondents, including those who are new to the SOEP. Importantly, most individuals who are currently exempt from mandatory insurance (such as civil servants and the self-employed) have one or more previous periods in their biography that were relevant to the pension insurance (e.g., periods of military service or enrollment in higher education). Since pension records contain this information, these data offer an enhancement to the SOEP data. Third, the linked data allow cross-validation of information in both datasets. Fourth, the administrative insurance biographies are set up as spell data, whereas the SOEP data (with the exception of the retrospective biographies) provide measurements at specific points during the year. SOEP-RV also expands the potential for research with the administrative data: The individual-level information in the administrative data complements the SOEP's detailed information on family and household relationships. In sum, SOEP-RV is especially useful for

describing and explaining the employment, pension, and income biographies of individuals and households. Furthermore, it allows the quantification of (lifetime) income—at the individual and household level—while taking into account earned and pension income as well as other types of income, including capital income and government benefits, without the need to rely on strong modelling assumptions. Analogous possibilities arise for the measurement of wealth according to asset types, including pension entitlements, as well as debts.

2 Linked Data Sources: SOEP, RTBN and VSKT

2.1 SOEP

The SOEP is an ongoing longitudinal survey of private households in Germany that has been running since 1984 (Goebel et al. 2019). Various refresher and supplementary samples have been added over time. Since 2010, the SOEP has surveyed more than 25,000 individuals annually. Participation in the survey is voluntary; nevertheless, the annual re-survey rates are very high, averaging about 94% over many years. SOEP's survey is interdisciplinary, covering a broad set of individual and household-level variables including socioeconomic status, political attitudes, psychological and health indicators, satisfaction and worries, expectations, family background, and education. Further, SOEP includes information on age, employment and retirement status, income types (including pensions), assets and debts. Overall, these variables provide a very detailed picture of employment and retirement histories at both the individual and household level, with extensive research potentials (Schröder et al. 2020).

2.2 RTBN

The RTBN is an administrative dataset containing all monthly pension payments paid out by German Pension Insurance in December of a given year.[1] Every observation represents one pension and distinguishes between old-age pensions and survivor or invalidity pensions.[2] For each pension, in addition to the amount, type, and exact starting point, the data include a range of important information, such as deductions

1 DRV Bund: https://statistik-rente.de/drv/extern/rente/documents/RTBN_Renten_nach_SGB_VI_und_sonstige_Renten_Gesamtueberblick.pdf [accessed on January 26, 2021].

2 In accordance with SGB VI, the RTBN includes all pension types. For SOEP-RV, the most relevant pension types are invalidity pensions, all types of old-age pensions (e.g., disability, old age, unemployment, (very) long-term insured) and survivor's pensions.

for early retirement or premiums for postponing retirement (Lüthen 2016).[3] The RTBN thus offers detailed information complementing the SOEP, allowing researcher's insight into questions such as precisely how and when individuals make the decision to retire, what deductions they were willing to accept, and whether the retirement decision was made due to poor health. Further, the data can proxy time of death and provide new avenues for mortality research. Last, the RTBN includes survivor pensions, which allows researchers to derive the lifetime income of the deceased partner. However, SOEP-RV cannot directly link data on survivors' pensions, although it collects information on the existence of such pensions if the deceased individual agreed to participate in SOEP-RV prior to his/her death.[4]

2.3 VSKT

To calculate pension entitlements, the German Pension Insurance carefully collects information on all contributors' earnings histories. The VSKT is the statistical image of these records. For each month between the ages of 14 and 67, the VSKT provides a monthly history covering employment, unemployment, sick leave, and earnings points, which are used to compute monthly gross earnings. Due to its biographical nature and monthly detail level, the original VSKT sample is frequently used in economic research, for instance, for studies on long-term inequality in lifetime earnings (Bönke et al. 2015) and for research on old age (e.g., Lüthen 2016). The biographical nature of the VSKT serves as a blueprint for SOEP-RV: If an individual gives consent to SOEP-RV, their biographies are retrieved from pension records in the VSKT format. This is even true for the already retired population. Therefore, SOEP-RV provides a unique possibility for analyzing the entire biographies of the resident population of Germany.

3 Consent, Selectivity, and Weighting

3.1 Consent and Selectivity

SOEP respondents were asked to consent to data linkage in 2018. Recently integrated new subsamples were exempted from this to reduce the risk of panel

3 For more information, see the code plan of RTBN 2018: http://forschung.deutsche-rentenversicherung.de/ FdzPortalWeb/getRessource.do?key=pufrtbn18xvsbb_cdpln.pdf.

4 The pension insurance stores survivors' pensions under the deceased person's social security number. Since we are unable to ask for consent here, we cannot retrieve the respective pensions.

attrition. In subsequent waves, the SOEP has made an effort to link these originally exempted individuals as well as SOEP respondents who were too young to give consent in 2018. However, until all of the SOEP samples have been asked for consent, the *SOEP population asked for consent* constitutes a *subsample of the overall SOEP adult population*. The *SOEP population asked for consent* makes up 14,966 respondents. Of those, 8,141 respondents (54.4%) gave consent and thus constitute the *consenting population*. This percentage of respondents consenting is in line with similar record linkage projects in Germany.[5]

The SOEP is equipped with survey weights that allow researchers to draw inferences about the base population: individuals living in non-institutionalized households in Germany. However, since the linked population is a subsample of SOEP's adult population, the question of selectivity naturally arises. We investigate selection with respect to observable characteristics in two steps: First, we use a multivariate logit model to investigate differences in the characteristics of the *adult population* and the *population asked for consent*. In the second step, we study differences between the *population asked for consent* and the *consenting population*. Our choice of explanatory variables in the multivariate models builds on evidence from comparable record linkage projects (e.g., Jenkins et al. 2006). In the following, we present the variables and briefly review exemplary previous evidence:

1. *Age*. Most studies show that consent decreases with age (Pascale 2011; Sakshaug et al. 2012a; Wahrendorf 2018; Weissman et al. 2016). In our case, closely related to age is *salience*. Here, individual knowledge about the nature of the linked data might influence consent. This implies that individuals who are about to retire may have different consent rates as they are well informed about their pension entitlements (Korbmacher and Schröder 2013).
2. *Health*. Physical limitations might negatively affect people's willingness to share their social security number (Jenkins et al. 2006).
3. *Gender*. Most studies find no gender-driven differences in the willingness to provide consent (Jenkins et al. 2006; Mostafa and Wiggins 2017).
4. *Migration background*. Migrants are usually found to be less likely to provide consent (Carter et al. 2010; Cruise et al. 2015; Sakshaug and Huber 2016).
5. *Place of residence*. Previous studies suggest differences between East and West Germany, with East Germans exhibiting higher consent rates (Antoni 2011; Coppola and Lamla 2012; Korbmacher and Schröder 2013).

5 SHARE-RV has a quota of 55% (http://www.share-project.org/special-data-sets/record-linkage-project/share-rv.html). SHARE-RV also links survey data to administrative pension data in Germany and constitutes the most comparable data research project.

6. *Education*. Mixed evidence: Whereas Carter et al. (2010) and Knies and Burton (2014) find a positive correlation, Kim et al. (2015) and Sakshaug et al. (2016) find negative effects. Others find different effects for particular levels of education or educational attainment (Jenkins et al. 2006)
7. *Income*. Mixed evidence: Some studies suggest a positive correlation (Carter et al. 2010; Huang et al. 2007; Mostafa and Wiggins 2017). Others find a higher consent rate for low incomes (Kim et al. 2015; Weissman et al. 2016), middle incomes (Coppola and Lamla, 2012), high incomes (Sakshaug et al. 2012b), or no relationship (Antoni 2011; Knies and Burton 2014; Korbmacher and Schröder 2013).
8. *Household composition*. Mixed evidence: Some studies indicate different consent rates across varying household compositions; others document different effects (Carter et al. 2010; Coppola and Lamla 2012).
9. *Homeownership*. Homeowners show repeatedly lower consent rates (Cruise et al. 2015; Yang et al. 2019).

As explained above, in no wave of SOEP-RV will all respondents in every possible subsample be asked for consent. The most important reason is that asking for consent potentially lowers the willingness of new SOEP respondents to participate. After individuals have taken part in several waves, enough trust has been established for SOEP to ask for consent to record linkage. It will therefore always be important to analyze who was asked for consent before analyzing the willingness to provide consent. Since SOEP survey weights are constructed for the entire SOEP, controlling for subsample participation by adjusting the survey weights helps in avoiding selectivity bias. This is especially true for the first waves of SOEP-RV: As this is the first time we have implemented the linkage procedure, the aim was to phase-in the linkage of further subsamples consecutively over time and focus on the oldest samples.

To explain statistically who in the adult SOEP population was asked for consent, we use logistic regression and show the results in terms of marginal effects in Table 1 (left two columns). We use the explanatory variables described above. Of course, depending on the research question, this list may need to be adapted, for example, when it comes to analyses by nationality. The reference group in the regressions is male respondents of age below 40 with a household post-government income in the bottom quintile, whose health and education is lower than medium; who have no migration background, and who are living in a 1-member household. In the first wave of SOEP-RV, we find that there was a higher probability of being asked for consent among older individuals and individuals with higher incomes. We also find higher probabilities for singles without children, people with medium education, and homeowners. The probability of being asked

for consent was lower for respondents with direct (first generation) or indirect (second/third generation) migration backgrounds. Furthermore, the probability of being asked for consent was lower for all types of household combinations in comparison to a single household. The results with respect to migration background, income, and household composition and are not surprising, as the first wave of SOEP-RV did not include most of the migration subsamples and only part of the subsamples of low-income families.

Table 1: Marginal effects after logistic regression: Who was asked for consent and who gave consent?

Variables	Who was asked for consent?		Who gave consent?	
	Margins	SE	Margins	SE
Age: 40–49	−0.010	(0.009)	−0.023	(0.015)
Age: 50–59	−0.021**	(0.009)	−0.019	(0.015)
Age: 60–69	0.073***	(0.011)	−0.026	(0.017)
Age: 70–79	0.111***	(0.013)	−0.080***	(0.018)
Age: 80+	0.169***	(0.017)	−0.089***	(0.022)
Medium health	−0.008	(0.008)	−0.013	(0.012)
Good health	−0.015*	(0.008)	0.019	(0.013)
Female	0.008	(0.006)	−0.002	(0.009)
Direct migration background	−0.313***	(0.007)	−0.080***	(0.015)
Indirect migration background	−0.178***	(0.013)	−0.078***	(0.025)
East German	−0.010	(0.007)	0.036***	(0.011)
Medium education	0.034***	(0.009)	−0.029*	(0.015)
High education	−0.021**	(0.010)	−0.047***	(0.017)
Second income quintile	0.082***	(0.010)	−0.007	(0.015)
Third income quintile	0.113***	(0.011)	−0.014	(0.016)
Fourth income quintile	0.156***	(0.012)	−0.020	(0.018)
Fifth income quintile	0.164***	(0.013)	−0.024	(0.019)
Couple without children	−0.053***	(0.010)	0.003	(0.015)
Single parent	−0.173***	(0.013)	−0.024	(0.023)
Couple with children	−0.114***	(0.011)	0.008	(0.017)
Multiple generation HH	−0.220***	(0.030)	−0.008	(0.055)
Other combination	−0.171***	(0.027)	0.117**	(0.051)
Homeowner	0.074***	(0.006)	−0.046***	(0.010)
Observations	23,975		12,869	
Pseudo R-squared	0.175		0.009	
Chi-square test	5,782		155.7	
Prob. > chi^2	0.000		0.000	

Own calculations based on SOEP.v36 and SOEP-RV.2018. Standard errors (SE) in parentheses. The base category for age is 18–39. The base category household-type is 1-member household. ***p<0.01, **p<0.05, *p<0.1.

To explain statistically who gave consent, we again make use of a logistic regression. Results in terms of marginal effects are also summarized in Table 1 (right two columns). The willingness to consent decreases in age, which is in line with evidence from other studies (e.g., Pascale 2011; Sakshaug et al. 2012a; Wahrendorf 2018; Weissman et al. 2016). We find no effects for health, gender, or income, which is in line with the ambiguous or zero effects often reported (e.g., Jenkins et al. 2006; Antoni 2011; Knies and Burton 2014; Korbmacher and Schröder 2013). Migrants and their offsprings are less willing to give consent, which constitutes a typical result (e.g., Carter et al. 2010; Cruise et al. 2015; Sakshaug and Huber 2016). Highly educated individuals are less likely to consent, which is in line with some studies (Carter et al. 2010; Knies and Burton, 2014). Last, in line with the literature, homeowners are less likely to give consent (Cruise et al. 2015; Yang et al. 2019). In sum, our results are in line with the overwhelming majority of the literature and further confirm the unanimous results of no substantial consent bias.

3.2 A Reweighting Procedure to Adjust for Selectivity

There are two potential sources of selection: Selectivity of the *SOEP population asked for consent* and the selectivity of the consent among those SOEP subjects who were asked. To adjust SOEP frequency weights accordingly, we propose a four-step procedure recommended in a comparable context in Siegers et al. (2020) [6]:

Step 1: Estimation of a logistic regression model for the overall *SOEP adult population* where the dependent variable is a dummy variable indicating whether respondents were asked for consent to linkage of their SOEP data with the administrative data (dummy is equal to one) or were not asked (dummy is zero).

Step 2: If at least one explanatory variable is significant (e.g., p-value below 0.05) and at the same time shows a meaningful quantitative effect, the model is re-estimated only including the significant variables, and a correction of the SOEP survey weights is performed by multiplying the survey weights by the inverse estimated probability.

Step 3: Estimation of a logistic regression model for the population asked for consent where the dependent variable is a dummy variable indicating whether

6 Siegers S, Steinhauer HW, Zinn S. Gewichtung der SOEP-CoV-Studie 2020, SOEP Survey Papers, Series C – Data Documentation. 2020; 888.

respondents consented to data linkage (dummy is equal to one) or not (dummy is zero) using the same explanatory variables as in step 1.

Step 4: If at least one explanatory variable is significant and at the same time shows a meaningful quantitative effect, the model is re-estimated only including the significant variables, and the adjusted weights from step 3 are multiplied by the inverse estimated probability.

These double-adjusted SOEP weights yield the adjusted weight that can be used to calculate population statistics.

4 Comparisons of Key Variables

4.1 Validity of Information for Linked Cases

To validate the linkage, we compare RTBN information to self-reported information for successfully linked SOEP respondents. This section also serves as a warning to read the variable descriptions in both data sources as certain differences lie in the nature of the datasets.

We compare—at the level of each linked individual—the information contained in the two datasets on gender, marital status, age, and monthly retirement payments and display the results in Table 2. Our results suggest a near perfect match for both gender and age, which supports both a successful linkage and a correct collection of age and gender in the survey and administrative data. Further, Table 2 displays that marital status information deviates for about one fifth of the sample.

A naïve interpretation would be to argue that administrative data must be valid and that the survey data therefore provides false information. However, the devil is in the detail. In the SOEP, respondents are asked to provide marital status every year. By contrast, the pension insurance asks about marital status only when an individual applies for rehabilitation.[7] Therefore, (a) not everyone is asked this question, and (b) this information corresponds to a certain point in time in a persons' life.

Next, we compare individual monthly retirement payments as reported in SOEP and RTBN. The results are illustrated in Figure 1 by means of a scatter plot

7 Rehabilitation comprises medical and occupational rehabilitation.

Table 2: Comparison of gender and marital status information in the RTBN and SOEP for successfully linked respondents.

Variables	Consistent information	Inconsistent information
Gender	99.95%	00.05%
Age	100.00%	00.00%
Marital status:		
Single, divorced, widowed	92.42%	7.58%
Married	87.03%	12.97%
Missings	96.63%	3.37%
Observations	2,108	2,108

Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018.

with the number of cases underlying a particular combination of reported pay being reflected by the size of the bullets. This comparison involves several potential sources of error: First, SOEP values correspond to annual 2017 values because in 2018, the SOEP asked about the annual retirement payments received in the previous year, whereas the RTBN information represents December 2018 values. In most cases, this is the cause of very minor differences. However, some individuals who entered retirement late in 2017 reported 12-month values in the SOEP despite receiving pensions for fewer months, causing outliers: Here, we excluded four observations with extremely large SOEP retirement payments (up to €15,000 per month). All four observations had in common that they entered retirement very late in 2017 and that their self-reported pension values, when used as annual values, correspond to their (much lower) pension values in the RTBN. One approach would be to adjust these values by treating them as 12-month values. A second would be to exclude SOEP respondents who entered retirement after 2017. For the purposes of the present overview, we chose the latter.

Finally, we exclude invalidity pensions in the RTBN. Since these pensions are not awarded on a permanent basis, individuals might leave the insurance between 2017 and 2018. Hence, this temporary pension may distort the results. Still, same-year comparisons in subsequent waves of SOEP-RV improve upon all those results, for instance, when the RTBN 2018 is comparable to the SOEP 2019. However, due to regularly occurring adjustments to the pension scheme, some minor deviations are likely to remain even after careful adjustments.

Despite the aforementioned shortcomings, Figure 1 indeed suggests that differences in retirement payments for most individuals are usually small. Many differences are close to zero, especially for lower pensions. Further, SOEP and

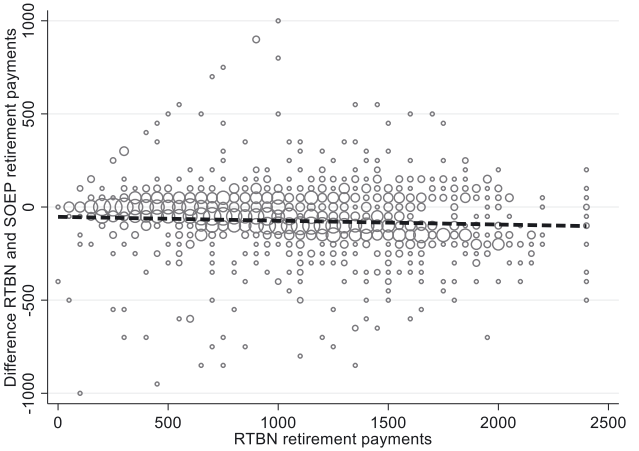

Figure 1: Differences in individual RTBN and SOEP monthly retirement payments.
Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The sample refers to 1,787 successfully linked SOEP respondents who entered retirement in 2017 or earlier and do not receive an invalidity pension. The difference refers to RTBN-SOEP monthly retirement payments.

RTBN retirement payments exhibit a positive correlation of 0.761. Nevertheless, a *t*-test of equal means suggests a €70 higher SOEP pension, which is significant (Table 3). Still, a *t*-test with bootstrapped percentiles further supports our results that deviations occur especially among individuals receiving retirement payments in the upper half. Or in other words: Individuals with retirement payments in the upper half report higher retirement payments in the SOEP than what is reported in the administrative data. This result is not surprising and actually underscores the advantages of the linked data: First, the RTBN censors monthly retirement payments of more than €2,199: Here the SOEP complements the RTBN and yields better information. Second, it is conceivable that the slight systematic upward deviation could be a result of older SOEP respondents who partially rounded up their retirement pay or mistakenly added other pensions such as widow pensions or company pensions.[8] In these cases, the RTBN delivers more precise information.

8 It cannot be ruled out that retirement income is no longer reported accurately due to the onset of dementia in old age.

Table 3: *T*-tests on equal means of the retirement payment variables.

Variables		RTBN	SOEP-RTBN	Difference	SE
Retirement payments	Mean	1,027	1,097	−70***	10.71
	P10	331	320	11	10.83
	P50	1,006	1,070	−64***	10.71
	P90	1,730	1,850	−120***	20.30
Observations		1,787	1,787		

Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN – SOEP-RTBN. The p-values result from *t*-tests of equal means and bootstrapped percentiles between the retirement payment variable of the RTBN and the retirement payment variable of the SOEP-RTBN for 1,787 successfully linked SOEP respondents who entered retirement in 2017 or earlier and do not receive an invalidity pension. SE are the standard errors. ***p<0.01, **p<0.05, *p<0.1.

4.2 Comparison of SOEP-RTBN and RTBN

To evaluate the overall representativeness of SOEP-RV for statutory pensions in Germany, we compare the RTBN variables gender, age, monthly retirement payments, and pension types for the linked SOEP-RV population to a representative 1% RTBN sample. We restrict both samples to individuals born in 1958 or earlier to ensure a proper comparison. Figure 2 shows that the SOEP-RV population is younger and receives higher pensions than the RTBN population. Further descriptive results in Table 4 confirm this pattern. Since the SOEP does not include individuals living in care facilities or comparable institutions, differences—especially for the very old—are to be expected.

For further insight, Table 5 shows a comparison by share of pension types. We find small and significant differences for some pension types, but the only larger deviation is found in regular old-age pensions (about 8 percentage points). Those differences stem from those older than 85, who predominantly receive old-age pensions and are underrepresented in the SOEP (Figure 2). Table 5 also reveals another strength of the record linkage: SOEP-RV allows investigation of the household situation of certain pension recipients, such as recipients of invalidity pensions, who experience a higher risk of poverty due to the interruptions in their employment histories.

5 Data Access

A central aim of this project is to make the resulting new dataset and its analysis potential available for scientific use as easily as possible and according to the FAIR criteria (Betancort et al. 2020). Both datasets—the SOEP survey data and the

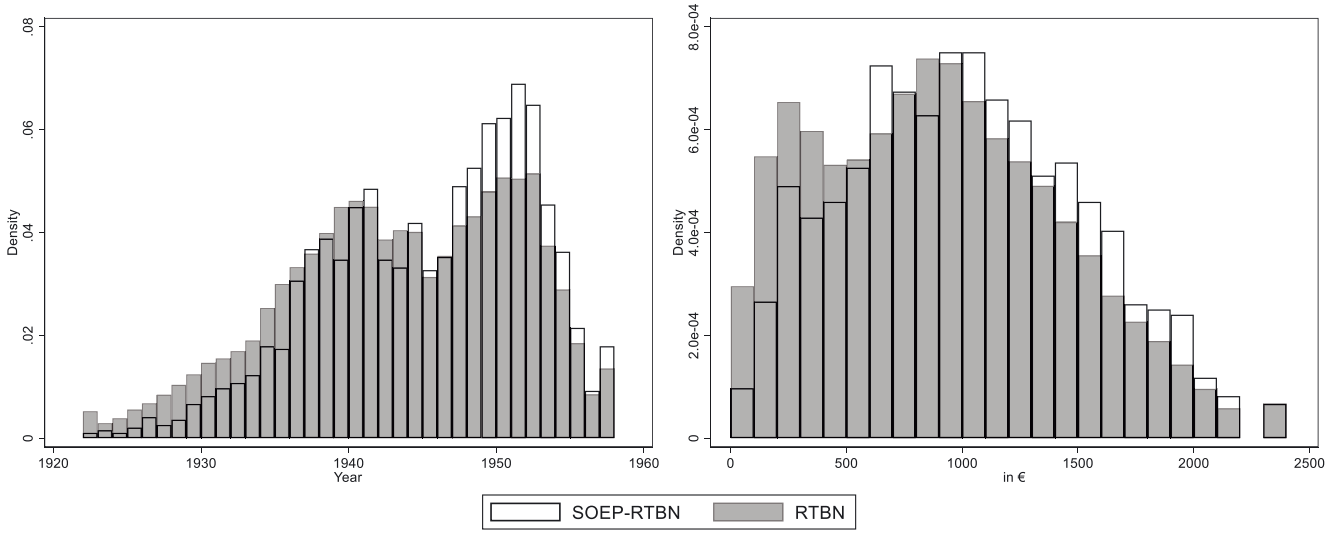

Figure 2: Birth year and retirement payments.
Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The samples refer to 1.958 successfully linked SOEP respondents in the SOEP-RTBN and a representative sample of 188,910 observations in the RTBN. Samples are restricted to individuals aged 60 or older.

Table 4: Descriptive statistics.

Variables		RTBN	SOEP-RTBN	Difference	SE
Age	Mean	75	73	2***	9.54
	P10	65	65	0	0.17
	P50	75	72	3***	0.62
	P90	86	83	3***	0.48
Retirement payments	Mean	904	1016	−112***	511.52
	P10	226	333	−107***	15.09
	P50	878	993	−115***	13.77
	P90	1,620	1,705	−85***	26.54
Observations		188,910	1,958		

Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN – SOEP-RTBN. The p-values result from *t*-tests of equal means and bootstrapped percentiles between the RTBN and the SOEP-RTBN samples restricted to individuals aged 60 or older. SE are the standard errors. ***p<0.01, **p<0.05, *p<0.1.

Table 5: Chi-squared test on equal proportions.

Variables	RTBN	SOEP-RTBN	Difference
Male	0.445	0.447	−0.002
Female	0.555	0.553	0.002
Invalidity pension	0.034	0.043	−0.009**
Regular old-age pension	0.404	0.325	0.080***
Unemployment/part time pension	0.108	0.112	−0.005
Old-age pension for women	0.189	0.199	−0.010
Pension for severely disabled	0.098	0.109	−0.011
Pension for long time insured	0.102	0.133	−0.030***
Pension for especially long time insured	0.063	0.080	−0.017***
Other pensions	0.002	0.000	0.001**
Observations	200,791	2,108	

Own unweighted calculations based on SOEP.v36 and SOEP-RV.2018. The difference refers to RTBN - SOEP-RTBN. The p-values result from chi-squared tests on equal means between the RTBN and the SOEP-RTBN samples restricted to age 60 or older. ***p<0.01, **p<0.05, *p<0.1.

administrative data of the German Pension Insurance—are available only for scientific research but free of charge. All datasets are provided for use in the statistical packages Stata and SPSS. Using the data for commercial purposes is forbidden. However, because of the different data sources (survey data versus social data), users must register separately at each Research Data Center according its access rules.

The SOEP survey data are available through the SOEP Research Data Center (RDC SOEP). After signing a data distribution contract, users can download data from all available years and subsamples with an individual download link. The link

is time-limited, encrypted, and can only be used in combination with a personal password, which is sent by text message to the user's cellphone. The administrative data are stored at and provided by the Research Data Centre of the German Pension Insurance (FDZ-RV). Data use requires registration and submission of an application form. After the registration process is completed, the data are sent to registered users on a hard disc. The final merging of the two data sources can be done by users themselves using the stable and unique identifiers included in both datasets.

6 Research Potentials and Concluding Remarks

We have provided an overview on the SOEP-RV-project, which connects SOEP survey data to administrative pension data through record linkage, offering many new avenues for research, especially on topics that require detailed pension information or long-term biographical employment and wage information on an individual or household level.

We have also documented that using the data is not as straightforward as it may seem. Because the SOEP has phased in the request for consent to data linkage starting with long-standing samples and asking newer samples only after trust has been established through participation in several waves of the survey, and due to the selectivity in consent, use of the SOEP-RV data requires weighting to be representative. To this end, we have illustrated an exemplary re-weighting procedure. We have also examined SOEP-RV data validity and explained differences in data from the two sources. Finally, we have explained how researchers can obtain the SOEP-RV data.

The project is still ongoing. Future data waves will open up even more avenues for research on topics such as mortality, and will include even greater numbers of individuals, improving representativeness.

Acknowledgments: We thank the Forschungsnetzwerk Alterssicherung (FNA), contract number 0640-FNA-P-2016-12, for financial support. We thank Deborah Anne Bowen for her careful editing of the paper. A more detailed version of the article including source codes can be found in the SOEP Working Paper Series.

References

Antoni, M. (2011). *Linking survey data with administrative employment data: The case of the German ALWA survey*. FDZ-Methodenreport, Nuremberg.

Betancort, C.N., Bongartz, E.C., Dörrenbächer, N., Goebel, J., Kaluza, H., and Siegers, P. (2020). White paper on implementing the FAIR principles for data in the social, behavioural, and economic sciences. RatSWDS Working Paper Series, https://doi.org/10.17620/02671.60.

Bönke, T., Giacomo, C., and Lüthen, H. (2015). Lifetime earnings inequality in Germany. J. Labor Econ. 33: 171–208.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Chapter 59: measurement error in survey data. In: Heckman, J., and Leamer, E. (Eds.), *Handbook of econometrics 5*. Chicago/London, pp. 3705–3843, https://doi.org/10.1016/s1573-4412(01)05012-7.

Carter, K., Shaw, C., Hayward, M., and Blakely, T. (2010). Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey. N.Z. J. Social Sci. Online 5, https://doi.org/10.1080/1177083x.2010.516440.

Coppola, M., Lamla, B. (2012), Empirical research on households saving and retirement security: first steps towards an innovative triple-linked-dataset. MEA Discussion Paper Series 201207.

Cruise, S.M., Patterson, L., Cardwell, C., and O'reilly, D.P. (2015). Large panel-survey data demonstrated country-level and ethnic minority variation in consent for health record linkage. J. Clin. Epidemiol. 68: 684–692.

Frick, J, R., and Grabka, M. (2005). Item-non-response on income questions in panel surveys: incidence, imputation and the impact on the income distribution. Allg. Stat. Arch. 89: 49–61.

Goebel, J., Grabka, M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2019). The German socio-economic panel (SOEP). Jahrb. Natl. Stat. 239: 345–360.

Huang, N., Shih, S.F., Chang, H.Y., and Chou, Y.J. (2007). Record linkage research and informed consent: who consents? BMC Health Serv. Res. 7.

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A., and Sala, E. (2006). Patterns of consent: evidence from a general household survey. J. R. Stat. Assoc. 169: 701–722.

Kim, J., Shin, H., Rosen, Z., Kang, J., Dykema, J., and Muenning, P. (2015). Trends and correlates of consenting to provide social security numbers: longitudinal findings from the general social survey (1993–2010). Field Methods 27: 348–362.

Knies, G., and Burton, J. (2014). Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys. BMC Med. Res. Methodol. 14: 14–125.

Korbmacher, J.M., and Schröder, M. (2013). Consent when linking survey data with administrative records: the role of the interviewer. Surv. Res. Methods 7: 115–131.

Lüthen, H. (2016). Rates of return and early retirement disincentives: evidence from a German pension reform. Ger. Econ. Rev. 17: 206–233.

Mostafa, T., and Wiggins, R.D. (2017). What influences respondents to behave consistently when asked to consent to health record linkage on repeat occasions? Int. J. Soc. Res. Methodol. 21: 119–134.

Pascale, J. (2011). *Requesting Consent to link survey Data to administrative records: Results from a split-ballot Experiment in the Survey of health Insurance and program participation (SHIPP)*. Survey Methodology, Washington.

Sakshaug, J.W., and Huber, M. (2016). An evaluation of panel nonresponse and linkage consent bias in a survey of employees in Germany. J. Surv. Stat. Methodol. 4: 71–93.

Sakshaug, J.W., and Kreuter, F. (2012a). Assessing the magnitude of non-consent biases in linked survey and administrative data. Surv. Res. Methods 6: 113–122.

Sakshaug, J.W., Couper, M.P., Ofstedal, M.B., and Weir, D.R. (2012b). Linking survey and administrative records: mechanisms of consent. Socio. Methods Res. 41: 535–569.

Schröder, C., König, J., Fedorets, A., Goebel, J., Grabka, M., Lüthen, H., Metzing, M., Schikora, F., and Liebig, S. (2020). The economic research potentials of the German socio-economic panel study. Ger. Econ. Rev. 21: 335–371.

Schnell, R. (2014). Linking surveys and administrative data. In: Engel, U.B., Jann, P., Lynn, A., and Scherpenzeel, S.P. (Eds.), *Improving surveys methods: Lessons from recent research*. New York: Routledge, Taylor & Francis Group, pp. 273–287, https://doi.org/10.4324/9781315756288-35.

Siegers, R., Steinhauer, H.W., and Zinn, S. (2020). *Gewichtung der SOEP-CoV-studie 2020, SOEP survey papers*, 888. Series C – Data Documentation, Berlin.

Wahrendorf, M., Marr, A., Antoni, M., Pesch, B., Jöckel, K.-H., Lunau, T., Moebus, S., Arendt, M., Brüning, T., Behrens, T., and Dragano, N. (2018). Agreement of self-reported and administrative data. Eur. J. Popul. 35: 329–346.

Weissman, J., Parker, J.D., Miller, D.M., Miller, E.A., and Gindi, R.M. (2016). The relationship between linkage refusal and selected health conditions of survey respondents. Surv. Pract. 9, https://doi.org/10.29115/sp-2016-0028.

Yang, D., Fricker, S., and Eltinge, J. (2019). Methods for exploratory assessment of consent-to-link in a household survey. J. Surv. Stat. Methodol. 7: 118–155.