Data Observer

Axel Börsch-Supan*, Christin Czaplicki, Sabine Friedel,
Imke Herold, Julie Korbmacher and Tatjana Mika

SHARE-RV: Linked Data to Study Aging in Germany

<https://doi.org/10.1515/jbnst-2018-0034>

1 Introduction

The project *SHARE-RV* generates new research data by linking the German Sample of the *Survey of Health, Ageing and Retirement in Europe (SHARE)* to selected administrative data of the *German Pension Insurance (RV)*. This article introduces the *SHARE-RV* data set as an instrument to study the aging process in Germany and its implications for individuals, households and society as a whole.

The research field of aging has become of widespread interest, especially in wealthy countries whose populations have a particularly large share of older individuals. These countries face many challenges since the changes in the population structure are rarely compatible with the structures of their often

***Corresponding author: Axel Börsch-Supan**, Munich Center for the Economics of Aging (MEA), Max-Planck-Institut für Sozialrecht und Sozialpolitik, Amalienstraße 33, München 80799, Germany, E-mail: Boersch-supan@mea.mpsoc.mpg.de

Christin Czaplicki, Deutsche Rentenversicherung Bund, Ruhrstraße 2, 10709 Berlin, Germany, E-mail: Christin.Czaplicki@drv-bund.de

Sabine Friedel, MEA, Technical University of Munich (Chair for the Economics of Aging), München, Germany; and Munich Center for the Economics of Aging (MEA), Max-Planck-Institut für Sozialrecht und Sozialpolitik, Amalienstraße 33, München 80799, Germany, E-mail: friedel@mea.mpsoc.mpg.de

Imke Herold, Munich Center for the Economics of Aging (MEA), Max-Planck-Institut für Sozialrecht und Sozialpolitik, Amalienstraße 33, München 80799, Germany, E-mail: herold@mea.mpsoc.mpg.de

Julie Korbmacher, BIDAQ - Bayerisches Institut für Daten, Analysen und Qualitätssicherung, Ringstraße 55, Haar 85540, Germany, E-mail: julie.korbmacher@kbo.de

Tatjana Mika, Deutsche Rentenversicherung Bund, Ruhrstraße 2, 10709 Berlin, Germany, E-mail: tatjana.mika@drv-bund.de

very generous welfare systems that were designed when populations were relatively young and growing. Although a lot is known about the main causes of demographic change – a lower fertility rate coinciding with an unbroken trend of higher life expectancy – there is still little knowledge about the complex interactions, e. g. between changes in health and changes in socio-economic status, the effects of longer working lives and the inequality in older ages that may arise from the changing structure of work due to digitalization. More research is needed to better understand how to deal with these trends. This is a challenging task since the combination of population aging and digitalization is a unique development in human history without precedence and a multi-dimensional phenomenon which affects almost all aspects of people's lives. Hence, complex data are required to explore this unique historical process.

Researchers can use basically two very different data sources: survey data and administrative data. Each has its respective merits. The multidisciplinary survey *SHARE* provides rich information about individual characteristics of people aged 50 and older, such as their health and education, plus information about the household context and partnership histories. Information linked from administrative data – especially very detailed employment histories – supplements this survey. The resulting data set *SHARE-RV* is therefore ideally suited to answer questions which involve interactions between the employment histories of individuals and their effects on socio-economic status and health at older ages in the context of the household in which they are living.

Since the establishment of “Research Data Centers” in Germany (Gramlich et al. 2010) administrative data are widely available for many research purposes. However, administrative data are produced in offices, agencies or companies as so-called by-products – thus not primary for research purposes. Still, advantages of administrative data are obvious: they usually cover the complete target population, they are much more accurate than survey data, and they can be provided with less financial costs and time investments. The information in the data is largely exact, there is only a low share of missing values, and data do not depend on the respondents' capability to remember certain facts (Calderwood and Lessof 2009; Kröger et al. 2011). An essential disadvantage of administrative data, however, is that there are only limited context variables which could serve to answer research questions. This is because administrative data are collected in a particular field of administration and therefore limited to what is needed in this particular context. A broad range of context variables – like information on education, family status, or the household – are always included in surveys, in which topics are mainly set by researchers themselves, who know about the importance of the social background of a person. However, scientific surveys are reliant on the respondents' honesty and willingness to provide information.

Thus, the combination of survey and administrative data is attractive, because the data sources complement each other. The project *SHARE-RV* provides – precisely through linkage of survey with administrative data – a research data set that offers a very comprehensive potential for scientific and socio-political analyses in the field of aging research. The acronym *SHARE-RV* stands for the combination of the German sample of the *SHARE* with administrative data from the *German Pension Insurance*. It is a cooperation project between the *Munich Center for the Economics of Aging (MEA)* as part of the *Max-Planck-Institute for Social Law and Social Policy*, which coordinates *SHARE*, and the *Research Data Center of the German Pension Insurance (FDZ-RV)* in Berlin, which provides the administrative data. The project started as a pilot study in 2009 as part of the third *SHARE* Wave and is currently running in the seventh *SHARE* Wave.¹

The remainder of this article is structured as follows. In section 2 we give an overview over the different data sets, followed by a description of the linked data set in section 3. In section 4 we have a closer look at the linkage procedure, before providing information on how to get access to the data in section 5 and section 6 concludes with information on the most recent data release and an outlook.

2 Data

2.1 The Survey of Health, Ageing and Retirement in Europe (SHARE)

The *Survey of Health, Ageing and Retirement in Europe*² is the largest pan-European social science panel study collecting bi-annually micro data on health, socioeconomic status, and social and family networks of people aged 50 and older. The first Wave of data collection was established in 2004 in eleven countries (Austria, Belgium, Switzerland, Germany, Denmark, Spain, France, Greece, Italy, the Netherlands, and Sweden). New countries joined in later waves to cover Europe's cultural, economic, and social diversity. By now *SHARE* consists of 26 countries of the EU as well as Switzerland and Israel. Data from England and

¹ We would like to thank the Volkswagenstiftung for the initial and the Forschungsnetzwerk Alterssicherung (FNA) for the follow up funding of the *SHARE-RV* project. We also gratefully acknowledge funding of the *SHARE* study from the European Commission (Horizon 2020), the US National Institute on Aging, and national sources, especially the German Federal Ministry of Education and Research.

² For a detailed description of the *SHARE* survey see Börsch-Supan et al. (2013).

Ireland are collected in the harmonized studies English Longitudinal Study of Ageing (ELSA) and the Irish Longitudinal Study on Ageing (TILDA).

The goal of *SHARE* is to cover the key areas of the older population’s lives which are economic circumstances, health, and social and family networks. The questionnaire is divided into several modules which contain a broad range of questions, measurements, and tests to collect detailed information on these key areas (see Table 1 for an overview). A couple of modules cover for example the socio-economic status with questions about labor force participation or retirement, income and pension claims, housing, assets, wealth, and consumption behavior. Moreover, information on expectations and well-being is gathered. A great strength of *SHARE* is also the availability of subjective and objective health measures including

Table 1: Overview of CAPI modules in SHARE (Wave 1, Wave 2, Wave 4, Wave 5, and Wave 6).

<i>Questionnaire Modules</i>	<i>Examples</i>
Coverscreen	Date of birth, gender, partner, household composition, interview date
Demographics	Education, marital status, country of birth & citizenship, parents & siblings
Physical Health	Self-rated health, diseases, weight & height, (I)ADL limitations [(instrumental) activities of daily living]
Behavioral Risks	Smoking & alcohol, nutrition, physical activity
Cognitive Function	Self- rated reading & writing skills, orientation, word list learning immediate & delayed recall, verbal fluency & numeracy
Mental Health	Hope, depression (EURO-D)
Health Care	Doctor visits, hospital stays, surgeries, forgone care, out of pocket payments
Employment and Pensions	Employment status, individual income sources (public benefits, pensions), job, work quality
Children	Number & demographics of children
Social Support	Help and care given and received
Financial Transfers	Money/gifts given and received
Housing	Owner (mortgages, loans & value), tenant (payments), type and features of building
Household Income	Income sources of all household members
Consumption	Expenditures for food, goods, services, ability to make ends meet
Assets	Bank and pension accounts, bonds, stock and funds, savings
Activities	Voluntary work, clubs, religious organizations, motivations, quality of life (CASP-12)
Expectations	Expected inheritances, life expectancy, future prospects
Interviewer Observations	Willingness to answer, understanding of questions, type of building, neighborhood

(continued)

Table 1: (continued)

Questionnaire Modules	Examples
<i>New modules after Wave 1</i>	
Since Wave 2: End-of-Life	Reasons and circumstances of death
Since Wave 4: Social Networks	Ego-centered network, contact, emotional closeness, geographical distance, satisfaction with network
Since Wave 5: Computer & Internet	Use of computer & Internet, self-rated skills
In Wave 5: Childhood	Health and relative position during respondent's childhood

Source: SHARE 2017.

physical activities, biomarkers, and tests to measure cognitive abilities. Social networks and social participation are measured by asking about living arrangements, activities, and received and given help. An important innovation is the social networks module which covers the composition (and change) of the respondent's networks.

To cover important life events before turning 50, *SHARE* collected data also retrospectively on life histories in the third Wave (*SHARELIFE*³). The topics are equivalent to the standard questionnaire including childhood living circumstances, family history (partners and children), the employment and housing history. The collection of life-histories was repeated in Wave 7⁴ for all respondents who did not answer that questionnaire before, such as new partners in the household, respondents from new countries that joined *SHARE*, and refreshment samples added after Wave 3.

In all countries, *SHARE* is based on probability samples of the population aged 50 years and older at the time of sampling. In addition to the target person, the partner – if living in the same household – is also asked to participate in the survey.

2.2 Administrative data

The research data center of the German Pension Insurance provides two types of administrative data which are of special interest to cover information of both parts of the employment history – the time period in which pension entitlements are accumulated and the period of the disbursement of these entitlements.

3 For more details on *SHARELIFE* see Schröder (2011).

4 *SHARE* Wave 7 data will be published in 2019.

The two data sets which can be linked to the German *SHARE* data⁵ are the ‘sample of the insured population’s records (VSKT)’ and the ‘policy holder pension portfolio (RTBN)’ which are drawn from the pension records.

The VSKT is a longitudinal data set with an additional cross-sectional part which covers people’s whole working history. This longitudinal data set is on a monthly base and contains the employment histories beginning with the age of 14 until the age of 67 (Himmelreicher and Stegmann 2008). For a sample in which the year of birth differs between 1919 and 1982 wrapping up over 90 years of employment history can cause difficulties due to changes in social legislation. Though, many social activities and situations across the life course do not differ in their social significance (e. g. receiving unemployment benefits), they differ in their legal preconditions and consequences for pension entitlements in a particular year. Because of the constant change of these legal conditions, the original information from the pension records would not allow the comparison over time. Therefore, the different situations have been recoded in one variable as 15 different kinds of ‘social situations’ which are rather stable over time. Out of these social situations paid employment has priority status and all other social situations are second in rank. That is: Only if no employment is registered, one of the other situations can apply. Besides the variable on social situations, all activities are additionally reported on separate timelines, so that overlapping or parallel activities besides paid employment can also be analyzed. Table 2 explores some examples of information which are available on a monthly base. For empirical analyses the data are offered to scientists in form of a

Table 2: Monthly information included in the VSKT.

Employment situation	Different states are listed such as education, employment, unemployment, and childcare
Care	Unprofessional care, meaning that people are not paid for that activity
Disability or sickness	Periods in which people cannot work due to disability or sickness
Unemployment	Periods of unemployment
Classification of the job	This variable differentiates between 15 different ‘social situations’
Children	Number of children who are younger than 36 month and number of children younger than 144 month
Earning points	These points allow to recalculate the salary and the value of pension contributions that result from a certain activity

⁵ Only if a respondent consented to the linkage, see 3. Data Linkage.

Scientific Use File (SUF). The SUF is produced by drawing a sample from the original pension records, leaving out information that could lead to personal identification such as the social security number, name and address, and the employer's name and address (Rehfeld and Mika 2006).

The cross-sectional part of the VSKT includes further demographic information from the latest recorded activity. This means that demographic variables mirror the social situation at the moment the data are drawn from the pension records. For retired persons the cross-sectional data are complemented by the policy holder pension portfolio data (RTBN). Included is information on the year and month of the first pension payment, the type of payment (as disability pension, old age pension, early retirement pension) and very detailed information on which activities had been relevant for the calculation of the payment as well as the amount of the payment. This information allows analyses of the legal proceedings and calculations of the pension insurance which led to the pension paid to the interviewed person (Mika et al. 2010).

3 The combination of *SHARE* and administrative data from the German Pension Insurance

The combination of the two data sources constitutes much more than the addition of its parts and offers some remarkable advantages over the usage of each single data set. Some examples are the following:

- 1) **Broad and specific:** The project *SHARE-RV* was designed in order to enable research on aging using information of a wide range of context variables on the one hand and very detailed and a very precise documentation of respondents' employment history on the other hand. The context variables include important life events, e. g. family events, health shocks, or changes in the networks as well as soft-facts like expectations and personality traits. A great strength of *SHARE* is also the collection of subjective and objective health measurements. The administrative data provide for example essential information about insured periods in the German Pension Insurance as well as details on the contributions paid within these periods. Combining the multi-disciplinary *SHARE* data with the longitudinal and detailed administrative data sets permits to go beyond classical retirement analysis by analyzing processes such as health influences on labor market decisions or pensions, as health plays a crucial role especially in the second half of life.
- 2) **Individuals and households:** There are many research questions for which it is useful not only to take individual determinants into account, but also

the household context of these individuals. This is especially important if the event of interest reflects a decision which is typically made in a couple setting (as for example the transition into retirement), but also for topics such as old-age poverty. *SHARE* covers the household context by also interviewing the partner who is living in the same household and by collecting information about the household composition. The administrative data are individual records which do not include information on partners or the household context. By linking the administrative data with *SHARE*, the household or partner information is integrated into the administrative data so that the whole employment biography of a couple can be analyzed in parallel. Accordingly, the advantages that accrue from the linkage are not limited to the individual level, but also hold true for the household level.

- 3) **The longitudinal dimension:** As aging is a process, the availability of longitudinal data bases is indispensable. Both data sources of *SHARE-RV* provide longitudinal information: The retrospective life histories of *SHARELIFE* cover the individual biographies beginning in respondents' childhood. These life histories are updated with bi-annual *SHARE* interviews with the same panel members that cover the current life circumstances and changes in these circumstances since the last interview. The administrative data are also longitudinal and cover the time period from age 14 until retirement. Consequently, there is a great overlap in the covered time period of the two data sets which allows detailed life-course analyses.
- 4) **Methodological research:** The overlap of information collected in *SHARE* and provided in the administrative records enables a validation of survey answers. This is of special interest for the data of *SHARELIFE* as there is only little knowledge about the quality of retrospectively collected data, since validation data are rare.

Altogether, *SHARE-RV* is unique in its composition and has tremendous advantages. Therefore, the question may arise why relatively few studies other than *SHARE-RV* link survey and administrative data. Besides limitations by data protection requirements, a possible answer can be the complex implementation and the elaborate data preparation process. In the next section we describe how this implementation and data protection are done in *SHARE-RV*.

4 Data linkage

How does *SHARE-RV* enable to link the German survey data of *SHARE* and the German administrative data from the German Pension Insurance? During the

SHARE interview it is asked for the respondents' Social Security Number (SSN), after the respondents have consented to the linkage. Then the survey data are linked to administrative data using the given SSN. This direct linkage procedure requires an explicit informed consent of the respondents which is collected on a consent form. The consent form has two main functions: First, it provides detailed information on the project *SHARE-RV* and the use of the personal data for research purposes. Second, it is a necessary legal formality to document the respondents' consent and collect the SSN.⁶ Before the data release the final anonymization step is to replace the SSN by the *SHARE* identifier that is fix across all waves.

All German *SHARE* respondents were asked for the first time for their consent to link their survey data with their administrative data in 2008/2009, in *SHARE*'s third Wave (Korbmacher and Czaplicki 2013). As consent is valid until the respondents' withdrawal, only new respondents or respondents who refused in previous waves were asked for consent in Waves 5 and 6. As a consequence, the administrative data sets include respondents who may have participated in different *SHARE* waves. That is to say, e. g. the administrative data of a respondent who consented in Wave 3 and took part only up to Wave 5 are still updated for the release of Wave 6.

The availability of administrative data which can be linked to the survey data, is summarized in wave-specific linkage rates (see Table 3). This rate is calculated as the number of respondents who consented and whose data can be linked,⁷ divided by all respondents who participated in that wave. The linkage rates refer to the most recent release version only. Since consent is also valid for previous waves, given consent effects the linkage rates of those, too. As the number of respondents differs from wave to wave due to attrition and new persons being interviewed, the linkage rate also differs across waves. Although the project *SHARE-RV* started in the third Wave, the linkage is not limited to data from this

Table 3: Linkage rate by wave.

	Wave 3 (Rel. 6–1-0)	Wave 4 (Rel. 6–1-0)	Wave 5 (Rel. 6–1-0)	Wave 6 (Rel. 6–1-0)
No. of interviews in <i>SHARE</i>	1921	1621	5752	4412
No. of linkable cases	1172	1105	4012	3358
Record linkage rate in %	61	68,2	69,7	76,1

⁶ For more detailed information please see Korbmacher and Schmidutz (2015).

⁷ Not all respondents who consented can also be linked. For example, these can be respondents who do not have an account at the German Pension Insurance or whose SSN is unknown.

wave on. All *SHARE* respondents have a stable identifier, which enables to link their administrative data with any prior wave if participation is given.

5 Data access

Both data sets - the *SHARE* survey data and the administrative data of the German Pension Insurance – are available free of charge for scientific purposes. Using the data for commercial use is explicitly not allowed. The data users have to register for both data sets separately, following the access rules of the respectively Research Data Centers.

The ***SHARE*** survey data are available via the *SHARE* Research Data Center⁸ after a successful registration as a user. The data of all available waves and countries can be downloaded centrally. The **administrative data** are stored at and provided by the Research Data Center of the German Pension Insurance.⁹ For that, an additional registration as well as an application form¹⁰ is necessary. Following a successful registration, the data are sent to registered users on a hard disc. Subsequently, the final merging of the two data sources via the stable identifier can be done by the users themselves.

All data sets are provided for the statistical packages Stata and SPSS. A detailed documentation can be found on the *SHARE* Homepage.

6 Outlook

Although the first waves of *SHARE* have significantly contributed to contemporary research on aging and have provided many insights, they have also left some questions open and raised many important new ones. To understand the processes of aging better, especially long-term panel data are needed. Hence, we plan to continue collecting *SHARE* data up to the tenth Wave. In Wave 7, life histories are collected from those respondents who became part of the *SHARE* panel after the third Wave (*SHARELIFE*). In Wave 8 an extended measurement of cognition will be implemented. In addition, German *SHARE* respondents, who

⁸ <http://www.share-eric.eu/data-access/user-registration.html>

⁹ <http://www.fdz-rv.de>

¹⁰ For detailed information on how to fill in the form, please check our User Guide (http://www.share-project.org/fileadmin/pdf_documentation/SHARE-RV_6-1-0_User_Guide.pdf).

have not consented yet, will be asked for their consent to record linkage. Waves 9 and 10 will focus on the retirement of the baby-boomers by including topics regarding the change in life before and after retirement.

In parallel, also the data linkage project *SHARE-RV* is planned to be continued until Wave 10. The aim is to provide a reliable longitudinal research data basis that combines the best of both worlds: survey and administrative data. We will keep working to increase the number of linked cases and to improve the user-friendly data handling features and documentation. *SHARE-RV* provides a unique combination of information that allows a variety of analyses. Especially in reference to the implications of pension reforms or adjustments, *SHARE-RV* offers a great potential. Two examples may make this point. First, regarding the appreciation of child-raising in the form of earning points, Mika and Czaplicki (2017) showed that even with the compensational earning points for child-raising, mothers have less earning points and consequently less pension than childless women. Second, a study by Börsch-Supan et al. (2015) had a closer look at the new early retirement pathway “retirement with 63”. They concluded that the original aim of this reform to give an early retirement option for underprivileged and perhaps ill individuals has not been achieved. The eligible workers for early retirement without deductions actually turned out to be healthier and had a higher net household income than the control group of ineligible workers.

In spite of the biennial collection of data, *SHARE* and *SHARE-RV* data are updated every year in spring. The regular data release ensures that the users can work with the most current version and with as many cases as possible. The number of cases available in the *SHARE-RV* data set increases whenever the number of cases in the administrative data increases. In addition, a new *SHARE-RV* data release includes the latest reporting year of the administrative data, and an extended match to *SHARE* data. At each new data release, the documentation is updated and utilities as well as support for users are extended. The current release version (*SHARE-RV* 6–1-0) includes the *SHARE* data until Wave 6, the VSKT, and the RTBN until 2015 and 2016, respectively. Overall, administrative data are available for almost 4400 *SHARE* respondents.

References

- Börsch-Supan, A., B. Alt, T. Bucher-Koenen (2015), Early Retirement for the Underprivileged? Using the Record-Linked *SHARE-RV* Data to Evaluate the Most Recent German Pension Reform. pp. 267–278 in: A. Börsch-Supan, T. Kneip, H. Litwin, M. Myck, G. Weber (Hrsg.), *Ageing in Europe - Supporting Policies for an Inclusive Society*. De Gruyter, Berlin.

- Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber (2013), Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*. Doi: 10.1093/ije/dyt088
- Calderwood, L., C. Lessof (2009), Enhancing Longitudinal Surveys by Linking to Administrative Data. pp. 55–72 in: P. Lynn (Ed.), *Methodology of Longitudinal Surveys*. Wiley, New York.
- Gramlich, T., T. Bachteler, B. Schimpl-Neimanns, R. Schnell (2010), Panelerhebungen Der Amtlichen Statistik Als Datenquellen Für Die Wirtschafts- Und Sozialwissenschaften. *ASTA Wirtschafts- Und Sozialstatistisches Archiv* 4 (3): 153–183.
- Himmelreicher, R.K., M. Stegmann (2008), New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV). *Schmollers Jahrbuch: Journal of Applied Social Science Studies/ Zeitschrift Für Wirtschafts-Und Sozialwissenschaften* 128 (4): 647–660.
- Korbmacher, J., C. Czaplicki (2013), Linking SHARE Survey Data with Administrative Records: First Experiences from SHARE-Germany. pp. 47–53 in: F. Malter, A. Börsch-Supan (Hrsg.), *SHARE Wave 4: Innovations & Methodology*. Munich Center for the Economics of Ageing, Munich.
- Korbmacher, J., D. Schmidutz (2015), A Note on Record Linkage. pp. 57–63 in: S. Share, F. Malter, A. Börsch-Supan (Hrsg.), *SHARE Wave 5: Innovations & Methodology*. Munich Center for the Economics of Ageing, Munich.
- Kröger, K., U. Fachinger, R.K. Himmelreicher (2011), Empirische Forschungsvorhaben Zur Alterssicherung. Einige kritische Anmerkungen zur aktuellen Datenlage. *RatSWD Working Paper Series* 170. Berlin.
- Mika, T., C. Czaplicki (2017), Fertility and Women's Old-Age Income. In S. Germany, M. Kreyenfeld, D. Konietzka (Hrsg.), *Childlessness in Europe: Contexts, Causes, and Consequences*. Springer International Publishing, Cham.
- Mika, T., U. Rehfeld, M. Stegmann (2010), Income Provisions and Retirement in Old Age. pp. 1107–1121 in: *Rat für Sozial- und Wirtschaftsdaten* (Hrsg.), *Building on Progress*. Opladen & Farmington Hills.
- Rehfeld, U., T. Mika (2006), European Data Watch: The Research Data Centre of the German Statutory Pension Insurance (FDZ-RV). *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift Für Wirtschafts-Und Sozialwissenschaften* 126: 121–127.
- Schröder, M. (2011), Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe SHARELIFE Methodology. Mannheim Research Institute for the Economics of Aging (MEA): Mannheim.