Deutsche Rentenversicherung

SHARE Survey of Health, Ageing and Retirement in Europe

Forschungs Daten Zentrum

SHARE-RV
User Guide

1. Introduction

SHARE-RV stands for the direct linkage of survey data of *the Survey of Health, Aging and Retirement in Europe* (SHARE: www.share-project.org) with administrative records of the German Pension Fund[1] (DRV: Deutsche **R**enten**v**ersicherung). SHARE-RV is a pilot study within the German subsample of the third wave of SHARE and part of the project "*A new perspective for aging research in Germany: linkages between disciplines (biology, medicine, economics, and social sciences) and linkages between data bases (socio-economic surveys, administrative records, and biomarkers*)" which is founded by the Volkswagen Foundation (VolkswagenStiftung) and consists of two main parts:

- linking survey data with innovative biomarkers and
- linking survey data with selected administrative records of the German Pension Fund.

This report applies to the latter part: the direct linkage of administrative and survey data.

Despite the fact that administrative data are not primarily generated for research purposes, they have many advantages compared to survey data. They often cover nearly 100% of the population of interest and they are more accurate then survey data because issues such as recall errors do not play a role (Calderwood and Lessof, 2009). However, administrative data are process data collected for a specific purpose (for example to calculate retirement requirements) and have their limitations with regard to content because they only consist of information needed for that specific purpose (Rehfeld, 2008). Surveys have the advantage that researchers can design the questions so that data needed to answer a specific research

[1] The German Pension Fund consists of 16 independent organizations. The procedure used here has been discussed with data privacy protection officers in all organizations. Furthermore, the project has been discussed in two councils of the self-government board of the public pension insurance. The positive decision of each data privacy protection officer and the two councils is a necessary condition for data linkage projects like SHARE-RV.

VolkswagenStiftung

Bundesministerium für Bildung und Forschung

SEVENTH FRAMEWORK PROGRAMME

question are collected. Here issues like unit or item nonresponse as well as recall error may reduce data quality and the external validity. Combining survey and administrative data thus is a fruitful way to combine the best of both worlds. The goal of the project SHARE-RV is to link survey data from SHARE with data from the Research Data Center of the German Pension Fund (in the following FDZ-RV) to provide researchers from different fields with a rich database.
In the following we describe the two data sources separately as well as the advantages and possibilities resulting from the linked data set. Paragraphs 3 and 4 describe data linkage techniques in general, the procedure utilized in SHARE, the verification process as well as problems and limitations of SHARE-RV. The last paragraph provides information about how to get access to the data and which data are available in detail.

2. Data overview

2.1 SHARE

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 45,000 individuals aged 50 or over. The three main pillars of SHARE are economics (income security and personal wealth), health (physical and mental health, disability, and mortality) and social networks (kinship and social networks, living arrangements; Börsch-Supan et al. 2005). SHARE respondents are interviewed every two years to monitor changes in these key life areas over time. Eleven countries contributed data to the 2004 SHARE baseline study. They are a balanced representation of the various regions in Europe, ranging from Scandinavia (Denmark and Sweden) through Central Europe (Austria, France, Germany, Switzerland, Belgium, and the Netherlands) to the Mediterranean (Spain, Italy and Greece). Further data were collected in 2005-06 in Israel and two 'new' EU member states – the Czech Republic and Poland – as well as in Ireland which joined SHARE in 2006. The survey's third wave, SHARELIFE, collects detailed retrospective life-histories in thirteen countries in 2008-09 with help of a life history calendar technique (Börsch-Supan et al., 2011).

2.2 Administrative data

To improve the data infrastructure between research and administrative statistics, several research data centers have been founded in Germany. One of them is the Research Data Center of the German Pension Fund (FDZ-RV) which was set-up in 2004 as an integral part of the German Pension Fund. The FDZ-RV supplies researchers with several cross-sectional and

longitudinal micro-data on topics like retirement, disability, and rehabilitation. The process-produced data are recorded for the purpose of administering the entitlements to the pension insurance and consist of data of all employees paying for social security contributions. This implies that not all Germans are included in the database; excluded are for example civil servants or freelancers. This affects only a small number of cases, however, since the majority of Germans (nearly 90%) do have a record (Mika, Rehfeld, and Stegmann, 2009).

The data provided by the FDZ-RV for research purposes are anonymised[2] subsamples drawn from the pool of individuals who are insured in the German Pension Fund. The dataset of the insured population (VSKT: Versichertenkontenstichprobe) is one of the longitudinal data sources of the FDZ-RV and includes insured persons and the state of their pension entitlements. For them detailed information on their insurance biography is available on a monthly basis beginning with the age of 15 until the maximum age of 67 (Himmelreicher and Stegmann, 2008). These biographies comprise basic demographic characteristics, the labour status (like periods of employment, care-giving for children, illness or unemployment) and finally the earning points for these activities (Stegmann, 2007).

To further enrich the data, cross-sectional pension data (RTBN: Versichertenrentenbestand) are included for retirees. The RTBN includes the amount of the pensions paid by the German Pension Fund and information about all entitlements used for the pension calculation. The longitudinal data set ends with the transition into retirement at the latest, but the additional pension data allow analysing respondents' lives also beyond their working period.

2.3 SHARE-RV

The combination of accurate administrative data and profound information about different aspects of the respondents' lives can provide a wide range of research possibilities. Since SHARELIFE covers a long episode of life, beginning in the early childhood until now, the temporal overlap between SHARELIFE and the administrative records is quite high. Additional to the enrichment of the SHARELIFE data by very detailed information on lifelong earnings, for example; the fact that some information is included in both data sets (like the job history) offers the possibility to validate the retrospectively collected SHARELIFE data.

Not only SHARE benefits from the linkage with another data base, linking also enriches the administrative records. Previous analyses of the labour biographies or the entitlements achieved and their influence on the old-age benefits lacked important contextual factors like information

[2] For further information on anonymisation procedures see Stegmann et al. 2005.

on the household or partnership. Furthermore, the wide range of topics measured in SHARE enable investigations on the connection between various aspects of the respondents´ lives and their working history or their socio-economic status in later life.

3. Technical aspects when linking survey and administrative data

3.1 Linking procedure and preparation of the data

The method of linking different data sources depends on legal and technical constraints of each data set. In a first step one has to define the type of data linkage:

- match data sources of the same person vs.
- match data sources of people who are similar (in a statistical sense).

Both procedures have advantages and disadvantages. When linking data from the same person, the respondents' consent is necessary (Calderwood and Lessof, 2009), which often decreases the number of linkable cases. When linking persons who are statistical similar, consent is not necessary but one has to deal with the fact that the linked data refers to a "statistical twin" only. The SHARE-RV project is based on a direct linkage procedure meaning that the records of exactly the same person (here SHARE respondents) were linked.

In a second step one has to decide *how* to link the different data sets to make sure that the correct data are used. There are two types of individual-level linkage: deterministic and probabilistic linkage (see Gill, 2001, for a detailed explanation, and Calderwood and Lessof, 2009, for an overview). The main difference lies in the fact whether or not disagreement between the matching variables is allowed. Deterministic matching (not allowing for disagreements) depends on a unique identifier available in both datasets whereas probabilistic matching (allowing for disagreements) is based on different linking variables which are all allocated with different weights.

SHARE-RV is based on deterministic matching using the Social Security Number (SSN) as a unique identifier for all Germans who have a record at the German Pension Fund. In contrast to other countries like the United States, the SSN in Germany is not a universal number which allows identification of citizens for the purposes of Social Security management or taxation (Bovbjerg, 2004), but an identifier for people who start working (or acquire social insurance contribution for other reasons).

The general structure of the SSN is statutorily defined in Article 147 of Book VI of the Social Code. The SSN comprises different information: At first it contains the area code, which

depends on the residence of the insured at the date the number is allocated. Furthermore, it includes the date of birth and the initial letter of the birth name, followed by a gender code of the insured person and an automatically generated check number. The format of the SSN allows checking this number via a comparison with the respondents' demographics and generating the number if it is completely missing. All elements of the SSN remain constant even if the residence, the workplace or the employer changes.

The decision to link data directly requires respondents' written consent within the SHARE interview. Given that the survey is computer assisted an additional paper form was necessary to collect the signature. Two steps of consent were necessary:

- Step 1: The first step was a verbal consent at the end of the SHARE interview. The respondents were asked for their consent to link the information just given in the interview with the administrative data held by the German Pension Fund.
- Step 2: If the respondents gave their consent, the interviewer handed out a consent letter that had to be filled out by the respondents themselves. This form recorded the respondents' SHARE-ID[3], the SSN, all information needed to generate and/or check the SSN as well as the signature. The respondents had to send this letter directly to the German Pension Fund so that neither the interviewer nor the survey agency would know the number. The SSN was only used to identify the respondents' DRV records and is not included in the resulting dataset.

Therefore, drop-outs are due to the fact that not all respondents were able or willing to fill in their SSN, some are not readable and some contain typos.

The FDZ-RV collected the arriving consent letters and gathered the information. To be absolutely sure that the correct records are linked, the German Pension Fund put a lot of effort in checking and correcting the SSN. In particular, the SSN and the information which is required to generate the SSN had to be examined. At first, it was necessary to ensure that the consent letter was signed as otherwise the FDZ-RV is not allowed to provide data. After collecting, documenting and checking the SSN (see paragraph 3.2 for a detailed description), the valid SSN were sent to the respective pension insurance institutions which manage the accounts. These institutions checked whether the SSN is available, extracted the data of the insurance accounts and dropped the SSN from the resulting data set. This step includes the transformation of the pension relevant records into statistical data. Subsequently the main demographics (gender, year and month of birth) were compared with the information given on the consent forms.

[3] The SHARE-ID is necessary to assign the consent form to a SHARE interview.

3.2 Verification procedures

The quality of the consent letters varies strongly: some letters were filled in completely whereas others suffered from missing information. In the correction process of the SSN two problems were identified:

- A first problem concerns the correctness of the SSN: one source of error is typos. As mentioned in paragraph 3.1 the SSN comprises demographic information, which can be used for an initial evaluation. Although some consent letters contained a SSN, the plausibility checks showed that some respondents filled in the SSN of their (deceased) partner for example.
- A second typical case is that respondents did not fill in their SSN but their personal data. However, if the personal data are available the FDZ-RV is able to generate the SSN.

Thus, it was possible to correct the SSN in nearly all cases where the form gives wrong or incomplete information.

To link the administrative data to the SHARE interviews, the SHARE-ID, which is also collected on the consent form, is used. Again, item-nonresponse or typos are possible sources of error. In order not to lose these cases for analysis, we used all information available in both datasets for the linkage. After linking all cases with a correct SHARE-ID, we linked the remaining cases via gender, month and year of birth. Only if all three variables were identical and only one match existed, the two spells were linked. If there was more than one case with exactly the same demographics, the information about the children was used in a second step. The number of children and years of child birth are available in both datasets, with one exception: in the records of the pension data children are listed for one parent because entitlements for children can only be collected by one of the parents. In the majority of cases this is the mother, so linking spells with help of information about children was mainly possible for women. These cases were linked if all demographics, the number of children as well as their birth years were identical.

In a last step, the demographics collected in SHARE and those included in the administrative records were compared automatically after all cases were linked. All corrections are documented in the additional data set “overview[4]” so that each user can decide whether or not he/she will include these cases in the analysis.

[4] “sharew3_rel1_gv_rv_link_overview.dta”

4. Problems and limitations

The most dominant factor responsible for a reduced number of linked cases is respondents` consent. But even if the respondent gives his/her verbal consent and sends the consent form back to the German Pension Fund, there are some cases where no administrative data are available. For those cases a record technically exists but the data cannot be provided by the German Pension Fund. There are some potential reasons for the missing data:

- Reimbursement of contributions: insured persons who paid contributions for a short period can ask for a reimbursement. As a consequence the data in their insurance accounts are deleted because all entitlements are compensated. For example, women who married between 1957 and 1967 were able to reimburse their contributions for periods of vocational training or occupation, which they accumulated before their marriage.
- Revaluation of former GDR pensions: Pensions including entitlements accumulated in the former German Democratic Republic have been calculated manually after the reunification. In these cases, the employment biography and other pension relevant events were not recorded by the German Pension Fund.
- Status of account clarification[5]: For example civil servants who paid contributions for a short period often clarify their pension accounts just shortly before their retirement; therefore these pension relevant times are not recorded. A special case is when the insured person paid social security contributions before the computerized acquisition was introduced. In these cases the contribution periods are not saved until the insured applies for account clarification.
- Decommissioning caused by double assignment of SSN: Although the SSN of an insured never changes, it is possible that a person has several SSN. This might happen if the insured person loses the social security card or an employer registers the employee when re-starting working by mistake. In these cases the institutions of the German Pension Fund allocate a new number whereby one of the pension insurance accounts is closed and those data are no longer available.

To capture the whole process of the linkage, all cases where a consent form was sent to the FDZ-RV (independent of the availability of data) are linked. Therefore the additional data set

[5] The account clarification is the process of recording all periods of pension relevant events to close possible information gaps in the insurance biographies.

"overview" consists of all cases of the actual release. The file includes information about each step of the linkage as well as a variable containing the reasons for missing data (see codebook).

5. Using SHARE-RV

5.1 Access to the data

To get access to the SHARE data, researchers have to register as SHARE users[6] in order to be able to download the SHARE data and the additional overview file in the SHARE Research Data Center[7]. Registered users will be informed automatically about new releases and other important news. The same procedure is necessary to get access to the administrative records of the FDZ-RV. These data are available for research purposes after a registration as a user via the website at the Research Data Center of the German Pension Fund, where a data application is available[8]. The data of the FDZ-RV will be sent on Compact Disk.

5.2 Linking the data

The project SHARE-RV was a pilot study within the third wave of SHARE (SHARELIFE). All respondents participating in this round of data collection were asked for consent. The linkage is not limited to data of the third wave but can also be linked with data of the same respondents of previous waves. The reference point for the description of the linkage is the German sample of SHARELIFE. Given that the number of cases can vary between the release versions, the total number of linkable cases depends on the release version. The descriptive linkage results are always based on the recent release version and will be updated with each release in the appendix of this report. Users have to check the release number of both data sets. For analyses the release numbers of the FDZ-RV data and the SHARELIFE data have to be the same.

As a result of SHARE-RV, different data sets are available. Initially, SHARE provides an overview for all cases of the recent release version including information about who gave consent, which data are available for each respondent and an indicator how the two data sets were linked (a codebook of this dataset is available on the website).

[6] (http://www.share-project.org/t3/share/fileadmin/pdf_documentation/SHARE_Data_Statement.pdf)

[7] Wave 3/Generated Variables/SHARE_RV Linkage Germany

[8] (http://forschung.deutsche-rentenversicherung.de/FdzPortalWeb/formular.jsp?form=antrag.jsp)

The administrative records are divided into two parts:

- VSKT: the biographies of the insured population[9] (also available for retirees)
- RTBN: the pension records (only available for retirees).

The two data sets are available from the DRV and include the mergeid (as the unique identifier for all SHARE-respondents across all waves) to link the data to all modules of SHARE.

PLEASE NOTE: When using SHARE-RV data for publications, please mention the VolkswagenStiftung as funder of the project SHARE-RV in the acknowledgments alongside the SHARE acknowledgement which has to be always included.

Responsible for preparation of the SHARE-RV data and the documentation are:

Julie Korbmacher
Christin Czaplicki

Version: 1.0 (20.12.2011)
Contact: info@share-project.org with “SHARE-RV” in the subject

[9] Users should note that the VSKT represents the insured population on the last day of the survey year. For those people who are not retired at this time, the concept of the VSKT assumes reduced earning capacities for the working population and thus the receipt of a disability pension from the first January of the following year onwards. For that reason the biographies of these people contain hypothetical values for the further life course based on the previous records. Given the large number of pensioners in SHARE the calculation of pensions by using the values of the VSKT would lead to a distortion of the pension level. Therefore we excluded values for the pension calculation from the VSKT. However, the VSKT still contains the pension entitlements for insured and not retired persons, which means that the investigation of the socio-economic situation for both insured and retirees is still possible.

6. Literature

Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J., Siegrist, J., Weber, G. (2005): Health, Ageing and Retirement in Europe - First Results from the Survey of Health, Ageing and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A., Schröder, M. (2011): Retrospective Data Collection in the Survey of Health Ageing and Retirement in Europe. In Schröder, M. (ed) Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe – SHARELIFE Methodology. Mannheim: Mannheim Research Institute for the Economics of Ageing (MEA).

Börsch-Supan, A., Brandt M., Hank K. and Schröder, M. (eds). (2011): The individual and the welfare state. Life histories in Europe. Heidelberg: Springer.

Bovbjerg, B. D. (2004): Social Security Numbers. Use Is Widespread and Protections Vary. United States General Accounting Office (GAO). GAO-04-768T. In: http://epic.org/privacy/ssn/d04768t.pdf (Opened 13.10.2011).

Calderwood, L., Lessof, C. (2009): Enhancing Longitudinal Surveys by Linking to Administrative Data, in Methodology of Longitudinal Surveys (ed P. Lynn), John Wiley & Sons, Ltd, Chichester, UK.

Gill, L. (2001): Methods for Automatic Record Matching and Linkage and their Use in National Statistics. National Statistics Methodology Series, No. 25. London: Her Majesty's Stationery Office. Retrieved Oktober 2011 from http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/index.html.

Himmelreicher, R. K., Stegmann, M. (2008): New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV). In: European Data Watch. Schmollers Jahrbuch 128 (2008) , 647-660. Duncker und Humblot, Berlin.

Stegmann, M., Luckert, H., Mika, T. (2005): Die Bereitstellung prozessproduzierter Daten der GRV im Forschungsdatenzentrum der Rentenversicherung (FDZ-RV). Grundsätze zur faktischen Anonymisierung von Mikrodaten und zu Gastwissenschaftler-Arbeitsplätzen. In: DRV 2–3/2005, S. 203–215.

Mika, T., Rehfeld, U. G., Stegmann, M. (2009): Provisions for old age. Income provisions and retirement. RatSWD Working Paper Series 112, Berlin: Council for Social and Economic Data (RatSWD).

Rehfeld, U. G., Mika, T. (2006): The Research Data Centre of the German Statutory Pension Fund. European Data Watch. Schmollers Jahrbuch 126 (2006), 121-127. Duncker und Humblot, Berlin.

Rehfeld, U. G. (2008): Daten und Fakten für die Rentenpolitik - von Geschäftsstatistiken zum Forschungsdatenzentrum der Rentenversicherung. In: Rolf, G.; Zwick, M.; Wagner, G.G. (eds.), Fortschritte der informationellen Infrastruktur in Deutschland. Festschrift für Johann Hahlen zum 65. Geburtstag und Hans-Jürgen Krupp zum 75. Geburtstag. Nomos Verlagsgesellschaft, Baden-Baden, 194-208.

Rehfeld, U. G. (2009): Das Forschungsdatenzentrum der Rentenversicherung in stetiger Fortentwicklung. In: DRV-Schriften Band 55/2009.

Stegmann, M. (2007): Aufbereitung der Sondererhebung "Versicherungskontenstichprobe (VSKT)" als Scientific Use File für das FDZ-RV. In: Die Versicherungskontenstichprobe als Scientific Use File. Workshop des Forschungsdatenzentrums der Rentenversicherung (FDZ-RV) am 30. und 31. Oktober 2007 in Würzburg, 17-33. DRV-Schriften Band 79.