Schmollers Jahrbuch 132 (2012), 443 – 451
Duncker & Humblot, Berlin

Biographical Data of Social Insurance Agencies in Germany – Improving the Content of Administrative Data*

By Daniela Hochfellner, Dana Müller, and Anja Wurdack

1. Introduction

The use of administrative data sources is getting more and more popular in various research topics. The majority of microeconomic evaluation studies in Europe, which is about 80 percent, are based on administrative data sources (Card / Kluve / Weber, 2010) and (Scioch / Oberschachtsiek, 2009). The main reason for using register based data is the multitude of information they provide. On the one hand they contain a high number of observed individuals, on the other hand precise information on characteristics like wages or employment. An advantage compared to survey data is that there is no existence of non-response or panel attrition (Kluve, 2006). Administrative data sources of the german social security system are generated mainly using two procedures: the notification of the social security and the internal processes of the respective agencies which collect the data.

In the project Biographical Data of Social Insurance Agencies in Germany (BASiD), assisted by the Federal Ministry of Education and Research, German administrative data on individuals of two social security agencies, namely the German Federal Employment Agency (BA) and the German Pension Insurance (GRV), were merged. The focus was to create an innovative adjusted dataset via the integration of multiple data sources and to improve the quality as well as the content of administrative data. The developed dataset is provided to the scientific community worldwide as a Scientific Use File as well as a weakly anonymous dataset accessible by on-site use.

* We would like to thank the Federal Ministry of Education and Research for funding the project as well our colleagues of the Research Data Centre of the German Pension Insurance and our colleagues of the Institute for Employment Research particularly Hans Dietrich, Steffen Grießemer, Peter Jacobebbinghaus, Elke Jahn, Steffen Kaimer, Claudia Lehnert, Joseph Sakshaug, Patrycja Scioch, Gesine Stephan, and Rüdiger Wapler.

The remainder of the article is designated to detail a description of the current version of the weakly anonymous BASiD data.[1]

2. The Social Security System in Germany

There are three Social Security Agencies in Germany which rely on the same notification procedure of the social security system in Germany, namely the Health Insurance, the GRV, and the BA. In the notification procedure the employer has to report several pieces of information on his employees liable to social security. You can distinguish between two kinds of stored information: Information which is necessary to compute the social security contributions and information which is solely collected for statistical purposes. In the first case it can be assumed that the quality of the administrative information is very high and precise. For the most part plausibility checks are executed during the notification procedure of the social security. In the second case the quality of the information is most likely to be lower. Plausibility checks are not implemented in general, because error checking can be seen as time consuming and therefore expensive (Wichert / Wilke, 2010). In the BASiD project it was possible to link data of GRV and BA.

2.1 The German Federal Employment Agency

The BA is the labour market's biggest service provider. The agency is responsible for the administration of the compulsory unemployment insurance, the calculations of the amount of benefit an unemployed individual receives and placement offer as well as consultation of the unemployed. The Institute for Employment Research (IAB), which is part of the BA, is allowed to generate and hold historical datasets out of these records.

2.2 The German Pension Insurance

A pension insurance account is obligatory for all employed persons in the private and public sector. Pension contributions mainly depend on earning points individuals get for their employment histories. For every employment notification of the social security system individuals get earning points dependent on their corresponding wage. Additionally, in the case of unemployment, pension contributions are paid out of the unemployment insurance. Consequently 90 percent of the German population is insured at the GRV. The advantage of the GRV's data is the account clarification. From the age of 30 on,

[1] The current version of the data set is *BASiD 5109*.

employees which are subject to social security get a regular information report, which contains the employment times that are relevant for annuity computation. This allows mistakes to be recognized and corrected (Richter / Himmelreicher, 2008).

3. Outline of the Data Sources

The BASiD-project combines several single data sources to get various information on the individuals of both institutions into one dataset. The different sources are the Sample of Insured Persons and their Insurance Accounts (VSKT) of the GRV as well as the Integrated Employment Biographies (IEB) and the Establishment History Panel (BHP) of the IAB (Figure 1). The data of the GRV are the basis for the linkage. The individuals drawn from the pension data were identified in the different data sources of the IAB.

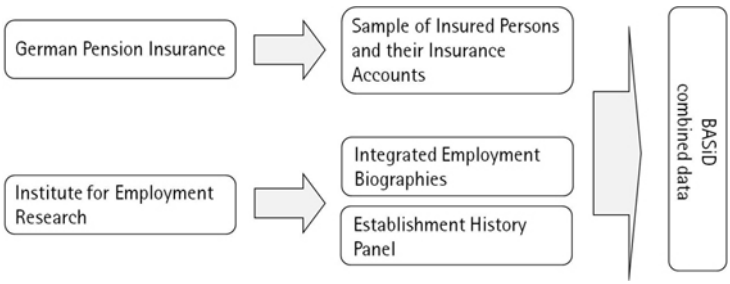

Figure 1: Administrative data basis of BASiD

3.1 Sample of Insured Persons and their Insurance Accounts

The VSKT is an annually generated sample of the GRV. It is drawn from all persons which notify at least one contribution in their insurance account at the end of each year. It provides information about every circumstance that is relevant for pension computations of the insured individuals. Therefore the biography of an insured person can be reconstructed at different points in time. Normally the recorded time span starts at the age of 17 and ends when a person gets her first pension payment (Himmelreicher / Stegmann, 2008). The information on the histories of the individuals is available since 1938 (Stegmann, 2008).

3.2 The Integrated Employment Biographies

The IEB includes information from different data sources. In the first place the data contains information about times of employment, which is stored in the form of a history dataset. It covers the time span from 1975 onward. Since

1st April 1999 notifications about marginal part-time employment are recorded additionally. These records regarding the employment history of individuals liable to social security are supplemented with information of the internal procedures of the BA. All notifications of the receipt of unemployment benefit, unemployment assistance or maintenance benefit since 1975 until the 1st January 2005 are added. On 1st January 2005 the receipt of unemployment assistance and maintenance benefit was pooled together and is called unemployment benefit II now. Additionally the dataset contains references to times of jobseeking and times of participation in active labour market policies (Jacobebbinghaus / Seth, 2007).

3.3 The Establishment History Panel

The BHP is generated by aggregation of single social security notifications to establishment level at 30th June each year. Therefore it contains every establishment in Germany that employs at least one person liable to social security or at least one marginal part-time employee since 1st January 1999 at that point in time. The BHP is constructed by yearly cross-sections since 1975. Since 1992 establishments in East Germany have been included (Spengler, 2009).

4. The Combined BASiD Data

The combined BASiD data differ in certain characteristics from previous existing datasets. It contains a variety of characteristics, which allow researchers to deal with research questions that could be answered for Germany only less precisely in the past. Another benefit is that the dataset contains complete employment biographies of individuals (Hochfellner / Voigt / Budzak / Steppich, 2010).

4.1 Content and Data Structure

The BASiD data contain longitudinal information on the life course of 568,468 individuals. It is arranged in an episode format. The covered period of time is from 1951 untill 2009. The observation period begins with entry into the education system and ends with entry into retirement. For instance, analyses with regard to birth-rates and employment histories of women, the influence of military or civil service on the employment histories, life-income and earnings points for the pension or influence of start-up-conditions on the career can be arranged. Figure 2 contains information from the different institutions. For a complete list of the variables in the BASiD data and a detailed description please refer to Hochfellner / Müller / Wurdack (2011).

	IAB	GRV
Employment and benefit history	X	X
Education (military and civil service)		X
Times of illness		X
Information on occupation	X	
Job seeking and training measures	X	
Job payments	X	X
Earning points and retirement characteristics		X
Motherhood and number of children		X
Regional and establishment information	X	
Sociodemographic information	X	X

Figure 2: Information in the final BASiD data

4.2 Sample Design

The basis for data fusion is the VSKT. The sample is a disproportionately stratified random sample by agency, gender, nationality and year of birth. Only persons between the age of 15 and 67 holding an account at the pension insurance on 31th of December 2007 for which contributions are made in the respective year were sampled. We searched for the same persons in the IAB data sets. Hence, the population of the BASiD data corresponds to all persons for which in 2007 contributions to pension insurance were recorded.

It is possible that not every included person necessarily holds an account at both institutions: For instance, if a person is self-employed but is voluntarily insured in the state pension system she can only be found in the data of the pension system. Due to the disproportionately stratified sample of social security numbers from pension accounts, representative analyses over time are difficult with the data set. For instance, women and insured foreigners or miners' have a higher sampling probability. However, there exists a weighting factor for the reporting year 2007, which compensates for the disproportionality and extrapolates to the population. In addition, the panel structure implicates that persons who are not alive on the reference day do not appear in the data.

5. Development of the BASiD Data

The preparation of the BASiD data was done in successively arranged steps: Data integration, data editing, comparison of simultaneous observations, and comparison with the Code of Social Law. In the following chapter we describe each step.

Administrative data of the BA as well as the GRV come from the same notification procedure of the social security system. Therefore we have an unique personal identifier available: the social security number. This identifier enables us to link the respective insurance accounts of the BA and GRV. But to find the corresponding notifications in every insurance account, we need additional identifiers. Therefore we used the start and end dates of the episodes, the actual state in employment history of the individual and the daily wage. As result unfortunately only 10 percent of the observations matched perfectly. The reason for the incomplete linkage is the different data management in each institution and data inconsistencies. Hence, the observations instead of being merged end up in an existence of multiple parallel spells in the linked data. To find the identical observations, what we call 'twin spell' within these multiple spells we developed strategies, which are discussed in the following.

To find more matches the identification variables had to be adjusted. At first we constructed identical observation periods via an episode splitting. Second we edited the daily wage because the format was different in each single data source. The calculation of the daily wage is based on working days as well as calendar days in the data sources of the BA, while the GRV uses calendar days continuously. Second the data of the GRV do not display the wage that is really earned by a person, but the wage that is relevant for the calculation of pensions.

After adjusting our identifiers we started the comparison of the simultaneous observations by the social security number, start and end dates of the episodes, the actual state in employment history of the individual, and the daily wage a second time. Identical observation periods between the BA and GRV were marked in the data and later transferred onto one single record. To document the searching routine we generated a variable which indicates the used searching routine. The searching routine is done by dividing the merged data in subsamples which are considered separately and illustrated in Figure 3.

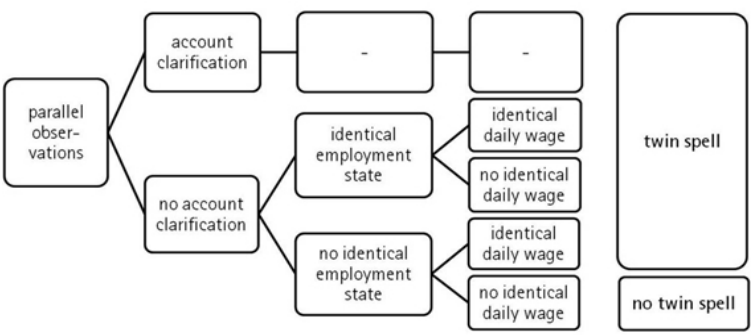

Figure 3: Comparison of simultaneous observations

We did not compare simultaneous observations if the insurance accounts were clarified. We assumed that information of the GRV are correct because there is a time-consuming clarification process of the GRV data (see page 444, chapter 2.2). For the simultaneous observations, which are not clarified, the employment state and the daily wage were compared. If the employment state and the daily wage were identical, we declared them as 'twin spell'. If the employment state was identical but the daily wage differed no more than one Euro we assumed a 'twin spell'. The daily wage was set to the mean of the different wage indications because it is not possible to decide which data source is more reliable. If the difference of the daily wage was larger than one Euro we considered the observations as 'twin spell' and the daily wage was set to missing. In case of no identical employment states the search for 'twin spell' was difficult. In the executed routine we differ between the attribute identical wage and no identical wage. If the wage was identical the employment state was corrected. If neither the employment state nor the wage were identical, we classified them as 'no twin spell'. A further comparison was executed on these. We compared 'no twin spells' with the code of social law to decide whether their parallel existence is possible or not. We kept them in the data as parallel observations if the parallel information is correct according the law. Otherwise we deleted the information. For a better understanding of the comparison of possible parallel episodes the example in Figure 4 and the description below illustrates the procedure in detail.

ID	SPELL	START	END	SOURCE	EMPLOYMENT STATE
1	1	01.01. 2000	31.12.2000	RV	Employment
1	2	01.01. 2000	31.12.2000	RV	Marginal part time employment
1	3	01.01. 2000	31.12.2000	RV	Maternity leave
1	4	01.01. 2000	31.12.2000	IAB	Job seeker
1	5	01.01. 2000	31.12.2000	RV	Pension payments

Figure: 4 Example for an existing no twin spell

In the merged dataset there may be a person that is employed. It is possible that this person has a second job at the same time, but only a marginal part-time job. Information concerning employment times is stored in the BA / IAB data as well the GRV data. Consequently for both observations there had to be found a twin spell in the corresponding data source in the first time. In this example there was 'no twin spell' found in the first searching routine because the employment episodes were missing in the BA data. Beside employment times there are allowance times in the GRV data like maternity leave. These can be seen as additional information to the respective employment relationship. Furthermore the employed person in our example is registered in the job

seeker register. A further employment state, which can exist at the same time, is an observation concerning pension payments which is valid only with restrains. In the displayed case this state is only valid when the pension payments are due to an entitlement to an orphan's pension.

After the comparison with the Code of Social Law most of the analysed sequences can be seen as conforming to the Code of Social Law.

6. Data Access

The FDZ BA / IAB currently offers access to the weak anonymised BASiD via on-site use at the FDZ BA / IAB and subsequent remote execution. Before data access is granted, an application form has to be completed by the researcher, approved by the Federal Ministry of Labour and Social Affairs (BMAS), and a contract with the FDZ BA / IAB has to be signed. Scientific use of social data requires the following conditions to be met and stated in the original request for data usage: scientific research regarding social security (§ 75 SGB X), prevailing public interest and permission of the BMAS (Dorner / Heining / Jacobebbinghaus / Seth, 2010).

The FDZ BA / IAB coordinates the whole application process of researchers. Specific application forms, guidance and further information on the different ways of data access can be found on our web page. Based on the data use agreement, researchers are provided direct on-site access at the FDZ BA / IAB in Nuremberg and other locations in Germany. On-site access is also possible in the United States at the Institute for Social Research at the University of Michigan. After a research visit at the FDZ BA/IAB, researchers can decide to continue data processing via remote data execution.[2]

7. Conclusion

The result of the successfully completed project is the combined BASiD data. The weakly anonymous version of the BASiD dataset is available since January 2012 at the FDZ BA / IAB. The combined BASiD data differ in certain characteristics from the previous existing datasets. It contains a variety of characteristics, which allows researchers to deal with questions that could be answered for Germany only less precisely in the past. An update of the dataset will take place if an increasing demand from the scientific community and a sustainable financing concept will emerge.

---

[2] Remote execution means that the researcher uses test data to prepare statistical codes and sends it to the FDZ BA / IAB by e-mail. The FDZ BA / IAB staff executes the codes, checks the generated output to avoid the identification of persons and sends the anonymised results back to the researcher.

References

Card, D. / *Kluve*, J. / *Weber*, A. (2010): Active Labor Market Policy Evaluations: A Meta-Analysis, The Economic Journal 120, 452 – 477.

Dorner, M. / *Heining*, J. / *Jacobebbinghaus*, P. / *Seth*, S. (2010): The Sample of Integrated Labour Market Biographies, Schmollers Jahrbuch / Journal of Applied Social Science Studies 130, 599 – 608.

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 (FDZRV), Schmollers Jahrbuch / Journal of Applied Social Science Studies 128, 647 – 660.

Hochfellner, D. / *Müller*, D. / *Wurdack*, A. (2011). BASiD – Biografiedaten ausgewählter Sozialversicherungsträger in Deutschland, FDZ Datenreportreport 09 / 2011.

Hochfellner, D. / *Voigt*, A. / *Budzak*, U. / *Steppich*, B. (2010): Das Projekt BASiD: Biographiedaten ausgewählter Sozialversicherungsträger in Deutschland. Projektinhalte, aktueller Stand der Arbeiten und Analysemöglichkeiten, DRV Schriften 55 / 2009, 74 – 86.

Jacobebbinghaus, P. / *Seth*, S. (2007): The German integrated employment biographies sample IEB, Schmollers Jahrbuch / Journal of Applied Social Science Studies 127, 335 – 342.

Kluve, J. (2006): The Eectiveness of European Active Labour Market Policy, IZA Discussion Papers, 2018.

Richter, M. / *Himmelreicher*, R. K. (2008): Die Versicherungskontenstichprobe als Datengrundlage für Analysen von Versicherungsbiografien unterschiedlicher Altersjahrgänge, DRV Schriften 79, 34 – 61.

Scioch, P. / *Oberschachtsiek*, D. (2009): Cleansing procedures for overlaps and inconsistencies in administrative data. The case of German labour market data, Historical Social Research 34, 242 – 259.

Spengler, A. (2009): The Establishment History Panel, Schmollers Jahrbuch / Journal of Applied Social Science Studies 129, 501 – 509.

Stegmann, M. (2008): Aufbereitung der Sondererhebung "Versichertenkontenstichprobe (VSKT)" als Scientific Use File für das FDZ-RV, DRV Schriften 79, 17 – 34.

Wichert, L. / *Wilke*, R. A. (2010): Which factors safeguard employment? An analysis with misclassified German register data, FDZ Methodenreport 11 / 2010.