Biographical Data of Social Insurance Agencies in Germany Improving the content of administrative data *

by Daniela Hochfellner, Dana Müller and Anja Wurdack

Abstract

The Research Data Centre of the German Federal Employment Agency in the Institute for Employment Research (FDZ BA/IAB) and the Research Data Centre of the German Pension Insurance (FDZ-RV) offer longitudinal individual-level data. These data contain information of the social security notifications as well as characteristics of the administrative procedures of both institutions. In each institution only information for accomplishment of their own current tasks is kept. The ambition of this project was to generate a unique dataset for the research community which contains data of both social security agencies. Therefore the richness of information on individuals was increased, through filling up gaps in the single data sources by using the information of the other data source. This linkage as well supports the improvement of the quality of administrative data.

1 Introduction

The use of administrative data sources is getting more and more popular in various research topics. For example, administrative data lend itself to policy evaluation or empirical labour market research in general. Nowadays

* 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, Patrycja Scioch, Gesine Stephan, and Rüdiger Wapler.

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 this is the multitude of information which is covered within these data. On the one hand they contain a high number of observed individuals, on the other hand they contain precise information on characteristics like wages, employment or unemployment periods and other social security transfers. An advantage compared to survey data is the fact that the existence of non-response or panel attrition is not given in register-based data (Kluve, 2006). Administrative data sources of the german social security system are generated mainly out of two proceedings, the notification of the social security and the internal processes of the respective agency which collects 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 new dataset allows researchers to accomplish more differentiated analysis in various research topics. The linkage leads to a unique employment biography dataset for researchers worldwide. The developed dataset is provided to the scientific community as Scientific Use File as well as a weakly anonymous dataset accessible by on-site use.

The remainder of the article is designated to detail a description of the current version of the weakly anonymous BASiD data [1] as well as to describe the framework of the project and the developed cleansing routines. It is organized in different subsections, which are structured as follows: In the first section the social security system and two of its Agencies are described shortly. The next section outlines the different databases which were linked. Afterwards, the new BASiD data are described in more detail. The following section displays the linkage and the developed cleansing routines. We conclude with a section addressing data access and resume the realized project.

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

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

the German Pension Insurance , the Health Insurance and the German Federal Employment Agency . In the notification procedure the employer has to report several 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 additionally 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). Unfortunately each institution only stores information for accomplishment of their own current tasks independent from the amount on information which the employer has to report. In the BASiD project it was possible to link data of two of the Social Security Agencies, which are described more detailed in the following.

2.1 The German Federal Employment Agency

The Federal Employment Agency is the labour market's biggest service provider. One area of responsibility is the administration of the compulsory unemployment insurance. The calculations of the amount of benefit an unemployed individual receives are drawn of the social security notifications. Another area of responsibility is the placement offer and the consultation of the unemployed. During these administrative processes a lot of records are generated. 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

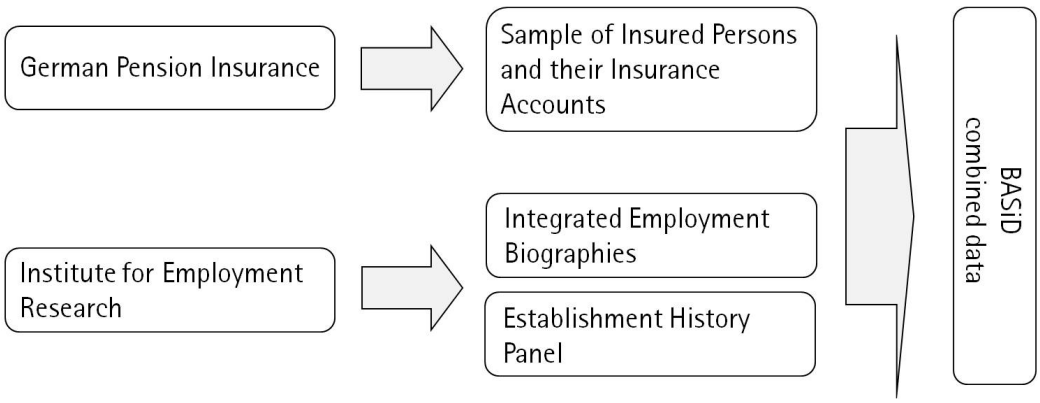
In Germany, a pension insurance account is obligatory for all employed persons in the private and public sector. Pension contributions mainly depend on earning points individuals get for their employment histories. For every employment notification of the social security system individuals get earning points dependent on the corresponding wage. The higher the achieved earning points, the higher the pension payments. Additionally, in the case of unemployment, pension contributions are paid out of the unemployment insurance. Hence individuals get little pension payments even for times in which they are not employed. Furthermore times of long term illnesses are

associated to pension contributions. During this time the health insurance is responsible for the payments. These illustrates that various states in life are taken into account for annuity computation. Consequently 90 percent of the German population have an account at the German Pension Insurance. The data of the Pension Insurance have one advantage, namely the account clarification, which can be used for the data cleansing. Account clarification means, that the Pension Insurance proofs the reported notifications. From the age of 30 on, employees which are subject to social security, get a regularly information writing, which contains the employment times that are relevant for annuity computation. This way originated mistakes are recognized and corrected (Richter & Himmelreicher, 2008).

3 Outline of the data sources

The project combined several single data sources to have various information on the individuals of both institutions in 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 Institute for Employment Research (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 Sample of Insured Persons and their Insurance Accounts (VSKT) is an annually generated sample of the German Pension Insurance. It is drawn

from all persons which at least notify 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. This means that the life situation of an insured person can be reconstructed at different points in time. Normally the recorded time span starts at the age of 17 and lasts until the date 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 Integrated Employment Biographies (IEB) include information of 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 on. 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 Federal Employment Agency. All deregistration 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 now called unemployment benefit II. Additionally the dataset contains references to times of job-seeking and times of participation in active labour market policies (Jacobebbinghaus & Seth, 2007).

3.3 The Establishment History Panel

The Establishment History Panel (BHP) is generated by the aggregation of the single social security notifications to the establishment level at 30th June each year. Therefore it contains every establishment in Germany that employs at least one person liable to social security at that point in time. Since the 1st January 1999 this is also true for establishments with at least one marginal part-time employee. The BHP is constructed by yearly cross-sections since 1975 in the case of establishments in West Germany and since 1992 for establishments in East Germany (Spengler, 2009).

4 The combined BASiD data

The combined BASiD data differ in certain characteristics from the previous existing datasets. It contains a variety of characteristics, which allows

the researchers to deal with research questions that could be answered for Germany only less precise 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 data contains all information gained from the social security notification process. This implies apprenticeship-, employment-, unemployment-, job-seeker-, training- and payment-details, pension times, consideration times, allowance times, payment dates and birth dates for children. Additionally the data contains establishment-information, like the establishment-size, classification of industries or regional information. For example 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. The subsequent table shows a basic classification and selection of the contained variables. For a complete list of the variables in the BASiD data and a detailed description please refer to Hochfellner, Müller, and Wurdack (2011).

Figure 2: Information in the final BASiD data

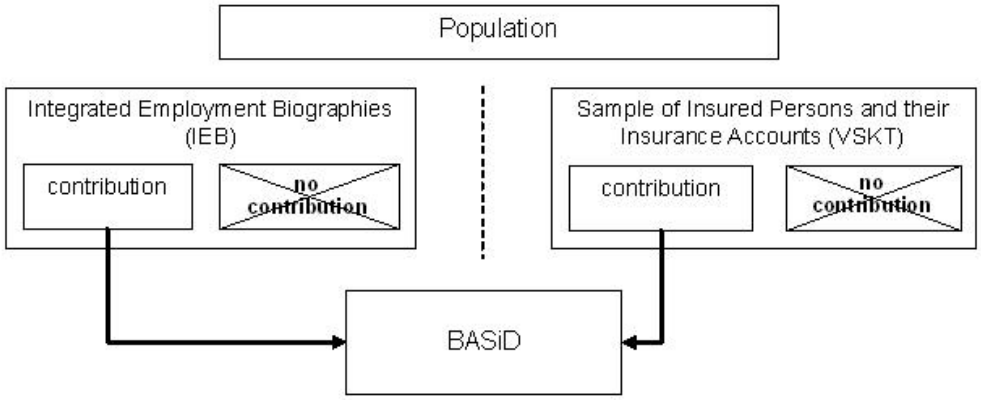
	IAB	GRV
Employment and benefit history	X	X
Education (military and civil service)		X
Times of illness		X
Information on occupation	X	
Job seeking and training measures	X	
Job payments	X	X
Earning points and retirement characteristics		X
Motherhood and number of children		X
Regional and establishment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

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 to 2009 for West Germany as well as Eastern Germany. The observation period starts with the entry into the education system and lasts until the entry into retirement.

4.2 Sample Design

Since the VSKT is the basis for data fusion, the design of this sample is as well the guideline for the BASiD data. The sample is a disproportionately stratified random sample by agency, gender, nationality and year of birth. From the master data of the RV, the gross sample is drawn. In a further step, only the accounts, which, at 31 December 2007, have an insurance account with the pension insurance, for which in 2007 contributes are paid, are transferred into the VSKT. The insured of the gross sample were ascertained in the IAB datasets, but only those from the VSKT were integrated in the BASiD data set. This population of the BASiD data thus corresponds to all persons, for which in 2007 contributions to pension insurance were recorded. Figure 4 points out the sample design.

Figure 3: Sampling design



Not every included person necessarily holds an account at both institutions: For example, if the person is self-employed but is voluntarily insured in the state pension system. By reason of this, the person can only be found in the data of the pension system. Due to the disproportionately stratified drawn social security numbers from the pension accounts, representative analysis over time not possible with the data set. For instance, women, foreigners or miners' insured have a higher sampling probability. However, there exists a weighting factor, which compensates for this disproportionality and extrapolates to the population. Since a weighting factor exists for the reporting year 2007, representative analysis are only possible for 2007. In addition, the panel structure implicates that persons, who are not alive at the current drawing time, do not appear in the data. They are replaced by a re-drawing.

A detailed description of the VSKT sampling design, which also applies to the BASiD data set, can be found at Richter and Himmelreicher (2008).

5 Development of the BASiD Data

The development of the BASiD data was done in successively arranged steps. The executed routines to merge the different data sources are described in the following chapters.

5.1 Data integration

Therefore the data of the Federal Employment Agency as well as the data of the Pension Insurance have a uniform identifier available: the social security number. Unfortunately this identifier is not enough for a successful data merging. The linkage of both data sources was done via the identifier, begin and end date of the episode, the actual state in the employment history of the individual and the daily wage. It can be assumed that there is no problem when merging the various data sources with the identifiers. Unfortunately there are still difficulties which occur during the linkage. In our case the information is stored in different formats in the multiple data sources. Furthermore we discovered inconsistencies which can lead to an existence of implausible multiple states in the final data. Due to this matter of facts it was not remarkable that only 10 percent of the observations match perfectly. Therefore we had to develop strategies to find the identical information in the different data sources.

5.2 Data cleansing

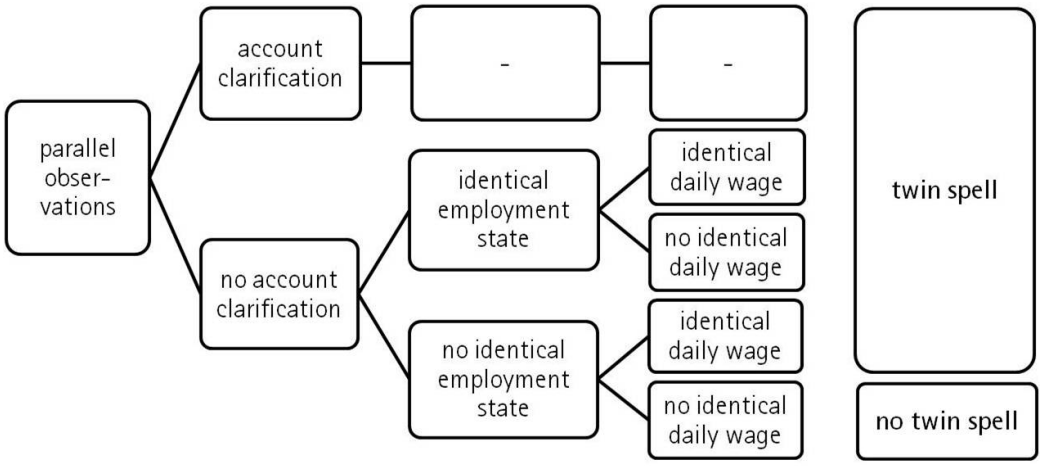
One of the inconsistencies can be explained through different observation periods. An episode splitting to construct identical observation periods was executed. After the episode splitting, the information in the data has to be adjusted to the new observation period. Another explanation why we could not find the identical observations during the data merging is that the information is stored in different formats in the single data sources. Because the daily wage is an identifier in the integration routine the format of this variable has to be adjusted, too. Three reasons apply why the daily wage differs between the two data sources. First of all the calculation of the daily wage is based on working days as well as calendar days in the data sources of the Federal Employment Agency, while the German Pension Insurance uses

calendar-days continuously. Second the data of the German Pension Insurance do not display the wage that is really earned by a person, but the wage that is relevant for the calculation of pensions. Because of that, the daily wages of employees in Eastern Germany have to be converted corresponding to annex 10 SGB VI § 256. Finally the adjustment of the currency change has to be executed.

5.3 Comparison of simultaneous observations

At this time of the data integration process all the identifiers we use to find identical observations (person, observation time, state in the employment history and daily wage), show the same structure and format. With the help of these identifiers identical information in the different data sources can be defined. In the following they are called "twin-spell". One of these can be equated with the existence of two identical observations. The searching routine is done by dividing the merged data in subsamples which are contemplated separately. The executed searching routine is described in figure 2.

Figure 4: Comparison of simultaneous observations



We did not compare simultaneous observations regarding the employment state and the daily wage if the insurance accounts are clarified. We assumed that the information of the Pension Insurance is the correct one for all considered combinations and did overwrite the corresponding information of the Federal Employment Agency. So, all observations that are clarified are recognized as "twin-spells". 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 the simultaneous observations as "twin-spells". If the employment state was identical but the daily wage differed we assumed a "twin-spell" when the deviation in the daily wage between both data sources is not larger than one. Because we cannot decide, which data source is more reliable, the daily wage was set to the mean of the different wage indications. If the difference of the daily wage was larger than one we considered the observations can be seen as "twin-spells" at least concerning the employment state and the daily wage was set to missing. More difficult was the searching for "twin-spells" in the case of no identical employment state. In the executed routine we differentiated between the attribute identical wage and no identical wage. If the wage was identical the employment state was corrected. The final decision for the correct employment state was made by comparing the information on the considered observations with the code of social law. Additionally plausibility checks within the dataset were executed. The only simultaneous observations which could not be identified as "twin-spell" differ in the employment state as well as in the daily wage.

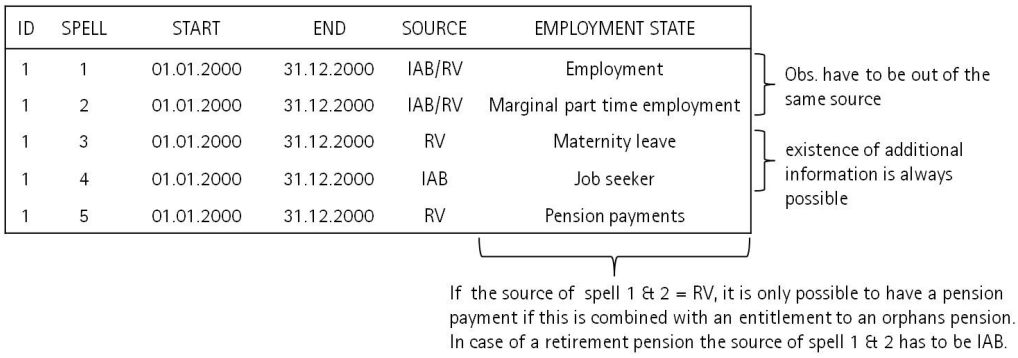
For each of these observations the information is compressed onto one single observation. In the developed routine the information on the observation of the German Pension Insurance has been transferred onto the observation of the Federal Employment Agency. After the transmission the duplicate is deleted. The transmission loop has to be executed several times, because the existence of multiple parallel episodes is possible.

5.4 Comparison with the Code of Social Law

For the existence of multiple employment states, the so called "no twin-spells" was proven, if their combination is possible. The checking was done by analysing sequences of different states that appear at the same time and comparing their existence with the Code of Social Law. Therefore a set of correspondence matrices was generated. Each matrix indicates for all simultaneous states in the employment history if their combination is possible regarding the German Code of Social Law. After analysing the German Code of Social Law we could distinguish if either the combination is valid or invalid. Exceptional cases are combinations which are only valid with restrains, e.g. the combination is only valid for a specific time because the Code of Social Law changed over time. Figure 3 illustrates the procedure in detail.

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, which only is

Figure 5: Example for an existing no-twin spell



a marginal part-time job. The information concerning employment times is stored in all data sources. 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. Therefore the observations have to be stored either only in the data of the German Pension Insurance or only in the data of the Federal Employment Agency. A mixture of both sources is not valid. Next to employment times there are allowance times, e.g. maternity leave stored in the data of the German Pension Insurance. These can be seen as additional information to the respective employment relationship and therefore are valid in every combination. Another state which is shown in figure 3 is called job seeker. The employed person in our example is registered in the job seeker register even though she is not unemployed. This situation is possible, because a person is not satisfied with her job and consequently registers as job seeker to find a new one. A further employment state, which can exist in 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. A regular pension payment can be rejected, because of the simultaneous maternity leave information.

After the comparison with the Code of Social Law most of the analysed sequences can be seen as conform regarding the Code of Social Law and will therefore stay in the data. Invalid combinations are corrected regarding the whole employment career of the individuals. The state in the employment history which does not fit into the biography was deleted.

6 Data access

The FDZ currently offers access to the weak anonymized BASiD via on-site use at the FDZ and subsequent remote execution. Before data access is granted, an application form has to be filled by the researcher, approved by the Federal Ministry of Labour and Social Affairs (BMAS), and a contract with the FDZ 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
- Permission of the Federal Ministry of Labour and Social Affairs (BMAS)

The FDZ 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 is provided direct on-site access at the FDZ in Nuremberg and other locations in Germany. You can also have on-site access in the United States at the Institute for Social Research at the University of Michigan. After a research visit at the FDZ, researchers can decide to continue data processing via remote data execution[2](Dorner, Heining, Jacobebbinghaus, & Seth, 2010).

7 Conclusion

The result of the successfully completed project BASiD is the combined BASiD data. The weakly anonymous version of the BASiD dataset is available since January 2012 at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (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 precise in the present. An update of the dataset will take place if an increasing demand of the scientific community and a sustainable financing concept will emerge.

[2]Remote execution means that the researcher uses the data documentation and the publicly available test data to prepare statistical code and sends it to the FDZ by e-mail. The FDZ staff executes the code and checks the generated output so that information suited to identify individuals is deleted from the statistical output. The remaining anonymised results are returned to the researcher by email.

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. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, *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 (FDZ-RV). *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, *128*, 647-660.

Hochfellner, D., Müller, D., & Wurdack, A. (2011). BASiD - Biografiedaten ausgewählter Sozialversicherungsträger in Deutschland. *FDZ Datenreportreport*, *09/2011*, 95 p.

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 IEBS. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, *2*, 45 p.

Kluve, J. (2006). The Effectiveness of European Active Labour Market Policy. *IZA Discussion Papers*, *2018*, 45 p.

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. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, *3*, 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*, 34 p.