Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

12/2015
EN

Methodological aspects of labour market data

Data protection aspects concerning the use of social or routine data

Stefanie March,
Angela Rauch,
Stefan Bender,
Peter Ihle

Bundesagentur für Arbeit

Data protection aspects concerning the use of social or routine data

Stefanie March (Institute for Social Medicine and Health Economics, Faculty of Medicine, Otto-von-Guericke University Magdeburg)

Angela Rauch (Institute for Employment Research)

Stefan Bender (Deutsche Bundesbank)

Peter Ihle (PMV Research Group at the Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital Cologne)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract 3

Zusammenfassung 3

1 Legal background 5
1.1 Overview 5
1.2 Social Code Book X (SGB X) – Transmission of social data for research 6

2 Organisational aspects 8
2.1 Data access: anonymisation versus pseudonymisation 9
2.2 Actors 11
2.3 Research data centres 12
2.4 Documents of relevance for data protection 13
2.5 Trust agency 14

3 Two examples of studies taking account of data protection legislation 15
3.1 lidA – leben in der Arbeit – study 15
3.2 Statutory Health Insurance Sample of Hesse 16

4 Conclusion 17

Abstract

When using social, respectively routine data in scientific research, extensive data protection rules have to be respected, particularly in the case of sensitive social data. This applies for both the use of microdata and the use of linked datasets (e.g. primary and secondary data). When planning and conducting scientific studies, extensive organisational and time resources are required. This article describes the legal and organisational requirements of data protection legislation in Germany. It explains terms such as social data, personal data, pseudonymisation, anonymisation and trust agency, and outlines the legal foundations. Key players/users, especially data owners, supervisory authorities and data protection officers, as well as their specific roles in the context of data protection are introduced. The key data protection steps are explained on the basis of different scenarios and selected examples. The article is intended to provide the reader with the necessary tools for using social / routine data for scientific research in compliance with data protection regulations beyond the German regulations.

Zusammenfassung

Bei der Nutzung von Sozial- bzw. Routinedaten im Rahmen wissenschaftlicher Forschung sind umfassende datenschutzrechtliche Vorgaben zu beachten. Nicht nur die Verwendung einzelner, sondern auch die Verknüpfung unterschiedlicher Datenquellen (z. B. Primär- und Sekundärdaten) muss datenschutzkonform erfolgen und bedingt einen umfassenden organisatorischen und zeitintensiven Aufwand, den es bei der Planung wissenschaftlicher Untersuchungen zu berücksichtigen gilt. Im Rahmen dieses Beitrages werden rechtliche und organisatorische Herausforderungen im Umgang mit den Vorgaben des Datenschutzes in Deutschland erörtert. Basierend auf den rechtlichen Regelungen werden Begriffe wie Sozialdaten, personenbezogene Daten, Pseudonymisierung, Anonymisierung oder Vertrauensstelle erläutert. Daneben werden die zentralen Akteure (Dateneigner, Aufsichtsbehörden, Datenschutzbeauftragte etc.) und deren Rolle im Kontext des Datenschutzes vorgestellt. Anhand verschiedener Szenarien und ausgewählter Beispiele werden die wichtigsten datenschutzrechtlichen Schritte erläutert, die bei der Planung von wissenschaftlichen Studien beachtet werden sollten. Dem Leser soll damit das notwendige Handwerkszeug für eine datenschutzkonforme Verwendung von Sozial- bzw. Routinedaten an die Hand gegeben werden.

Keywords: data protection, pseudonymisation, anonymisation, social data, administrative data, routine data

The original paper is - in a slightly amended form - published by the Hans Huber Verlag in 2014 (Reference: March S, Rauch A, Bender S, Ihle P (2014): Datenschutzrechtliche Aspekte bei der Nutzung von Routinedaten. In: Swart E, Ihle P, Gothe H, Matusiewicz D (Hrsg.): Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 2., vollst. überarb. Aufl. Verlag Hans Huber, Bern: 291-303). The authors thank the Hans Huber Verlag for given permission to publish this paper in an English version. All direct quotations in this paper (in quotation marks or in italic font) are translated from German into English. The content of the paper represents the authors‘ personal opinions and do not necessarily reflect the views of the Deutsche Bundesbank or its staff. All errors are our own.

1 Legal background

When planning and conducting secondary analyses with routine data or social data, data protection is an essential element of every research project. For this reason the legal basis, in particular regarding the transmission of data for a research project, for Germany is explained in detail in this paper.

1.1 Overview

General regulations concerning data protection are set down in the German Federal Data Protection Act (Bundesdatenschutzgesetz - BDSG) and in the data protection laws of the individual German federal states. In Germany specific regulations for the use of social data (hereafter also called 'routine data' or 'administrative data') can be found above all in Social Code Book X – Social and Administrative Procedures and Protection of Social Data (Zehntes Buch Sozialgesetzbuch – Sozialverwaltungsverfahren und Sozialdatenschutz - SGB X) as well as in other Social Code Books, in the Protection against Infection Act (Infektionsschutzgesetz) etc. The laws at federal state level may differ from state to state (AGENS 2012).

The legislator understands social data as "detailed personal or factual information about an identified or identifiable natural person (data subject)" (§ 67 (1) SGB X) that are collected, processed or used by a provider of social benefits (authorities, establishments or corporate bodies under public law and others,) or by an equal institution (see §35, SGB I) in the area of its responsibility.

In addition, "all firm-related or business-related data, also of legal entities, which are of a confidential nature" ((§ 67 (1) SGB X) are regarded as social data. Social data are subject to special data protection, personal data are highlighted as being especially sensitive; these include "details about racial and ethnic origin, political opinions, religious or philosophical beliefs, union membership, health or sexual life" (§ 67 (12) SGB X). Furthermore, according to the German Criminal Code (Strafgesetzbuch – StGB), medical data constitute a "secret which belongs to the sphere of personal privacy" (§ 203 StGB).

Moreover, in the European Union (EU), data protection at European level is regulated by Directive[1] 95/46/EG of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. In this directive, medical data are also granted special legal protection. In Article 8, for example, which concerns the processing of special categories of personal data, the Member States prohibit in Section 1 the processing of data concerning health. There are a few special cases in which this restriction does not apply, for example

> *processing is carried out in the course of its legitimate activities with appropriate guarantees by a foundation, association or any other non-profit-seeking body with a political, philosophical, religious or trade-union aim and on condition that the pro-*

[1] The definition of a directive in the European context and the importance of such directives for national law are regulated in detail in Art. 288 § 3 of the Treaty on the Functioning of the European Union.

cessing relates solely to the members of the body or to persons who have regular contact with it in connection with its purposes and that the data are not disclosed to a third party without the consent of the data subjects. (Article 8 (2/d))

Furthermore, the "Member States shall determine the conditions under which a national identification number or any other identifier of general application may be processed" (Directive 95/46/EG, Article 8 (7)). A comprehensive data protection act at European Union level is currently being developed but was not yet available in its final version when this paper was completed (reference date: 25.10.2015).

In addition, in Germany the principle of data privacy in the field of social security (Sozialgeheimnis) applies (§ 35 SGB I). This cites, among other things, the so-called prohibition with an authorisation proviso. In positive terms this means that everyone has the right that social data pertaining to his person shall only be collected, stored, processed (used) and transmitted if a legal basis exists. This basis can be found in the second chapter of Social Code Book X (SGB X). In § 67 (7) SGB X the legislator defines the use of social data as "any use (...), including the transmission of the data within the authority or office responsible". For research using routine data, the principle of data reduction and data economy also applies (§ 3a BDSG), i.e. the amount of data used must be restricted to that which is absolutely necessary for the research question.

Further guidelines for the different pools of social data can be found in the respective pertinent books of the Social Code. For instance, Social Code Book V (SGB V), Statutory Health Insurance regulates access to statutory health insurance data, e.g. in §§ 287, 303, 305. Depending on the research question or the pool of social data, further statutory regulations, also outside the sphere of SGB V, are to be clarified in advance.

As a consequence, these regulations mean that extensive data protection requirements have to be taken into account when social data are used for research. Anyone using social data for their research has to meet the challenges of different interest groups and try to satisfy them, especially in the sense of data protection. "Wherever personal data are evaluated, it takes place within the area of conflict between the right to informational self-determination that is laid down in Art. 2 (1) of the Basic Law for Germany (Grundgesetz – GG) (...) and the right to academic freedom (Art. 5 (3) GG). Finding an acceptable solution that takes both positions into account is therefore inevitably a balancing act between these poles (…) (von Ferber 1997). Ever since routine data began to be used for research, it has therefore always been necessary to take data protection aspects into account." (Ihle 2005, p. 195).

1.2 Social Code Book X (SGB X) – Transmission of social data for research

The transmission of social data for research purposes is regulated in Germany by the legislator in Social Code Book X (SGB X). In the context of this Code, compliance with various criteria is obligatory (§ 67 SGB X):

- The data subject must consent to the use of their data (§ 67b (1) sentence 1 SGB X).

- The data subject must be informed by the researchers about the purpose of the data use and about the consequences of refusing to provide consent (§ 67b (2) sentence 1 SGB X).
- The data subject must make the decision of their own free will (§ 67b (2) sentence 2 SGB X).
- The declaration of consent and the information regarding the research project must be in written form (§ 67b (2) sentence 3 SGB X).
- If, however, obtaining the data subjects' consent in writing proves to be unreasonable and inevitably compromises the research purpose, it may be dispensed with. The reasons are to be documented accordingly (§ 67b (3) SGB X).

In § 75 SGB X the special conditions are formulated under which it is permissible to transmit these data in the context of research projects and how this is to be done. Initially contrary to the stipulations set out in § 67 SGB X, it allows the transmission of data for research purposes even without the data subjects' consent (§ 75 (1) SGB X). However, certain requirements must be met: the data must be essential for the research purposes, i.e. the available social data must be the only possible source of the relevant information, and the public interest in the research must outweigh to a considerable extent the data subjects' private interest in maintaining the confidentiality of their data (ibid).

Furthermore, the transmission of social data requires prior authorization by the respective highest Federal or regional supervisory authority (§ 75 (2) SGB X). For this, the data owner generally has to submit a corresponding request in accordance with § 75 SGB X to his supervisory authority for permission to make social data accessible for research purposes. The request includes the following information among other things (ibid.):

> *1. the third party to which the data are transmitted,*
>
> *2. the type of social data to be transmitted and the group of data subjects concerned,*
>
> *3. the content of the scientific research [a description of the specific research project] or the planning for which the transmitted social data may be used, and*
>
> *4. the end date of the project until the transmitted social data have to be deleted (...).*

At the time when this paper went to press, it was still disputed between some actors whether a request for permission to transmit data in accordance with § 75 SGB X should also have to be submitted to the relevant supervisory authority when written consent is obtained from the data subjects (John und Krauth 2005, March et al. 2012).

It must also be taken into account that the legislator emphases "scientific research in the field of social benefits" (§ 75 (1) sentence 1 SGB X).[2] The use of the data for commercial purpos-

[2] This means that the social data may only be made available for research when the research is to be conducted in the same field as that in which the provider of social benefits that collects the data is active. In the case of the data collected and made available by the Federal Employment Agency

es is strictly precluded. Further interpretations regarding § 75 SGB X can be found in the annual report of the German Federal Commissioner for Data Protection and Freedom of Information (Bundesbeauftragter für den Datenschutz und die Informationsfreiheit - BfDI) (BfDI 2013).

The request for permission to transmit data in accordance with § 75 SGB X eases the burden of access to social data considerably for the research community, as a key methodological precondition for research is an undistorted reflection of the population. If, for example, a research institute is commissioned to conduct a survey and is given address details originating from the administrative procedures of a public authority, theoretically two selection steps would have to take place as described:

- Step 1: obtaining the potential respondents' permission to transmit the addresses and
- Step 2: obtaining a declaration of willingness to participate before the survey itself.

However, a problem of uncontrollable self-selection would arise here. A strong self-selection among the respondents can already be assumed to exist in Step 1 as this step demands a certain basic interest and activity of the potential respondents (e.g. Schnell 1997). Studies show, for instance, that an educational bias occurs in surveys. The younger, the more highly educated and the more satisfied a person is and the better their occupational position is, the more willing they are to participate (Lüdtke et al. 2003).

The consequence of self-selection is that every sample "recruited" in this way represents a systematically biased selection from the population. In order to avoid this, authorisation must always be obtained from the authority responsible by means of a request for permission to transmit data in accordance with § 75 SGB X whenever addresses from data stocks are to be transmitted without the respondents being asked to give their consent. Obtaining a declaration of willingness to participate before the start of the survey, Step 2, is imperative, however. Ideally, the potential respondents are informed about the planned research project in advance. This so-called participant information must also be accompanied, for example, by a data protection statement and information regarding the confidentiality of the data. Before the start of the survey, the respondents are then also asked whether they agree to participate in an interview. As a result of this prior information and due to the fact that the potential respondents do not need to be active themselves, the problem of self-selection is reduced as far as possible. This is a "low-threshold intervention" concerning the respondents and therefore every potential respondent should be given sufficient information in order to be able to decide whether to participate in the survey or to refuse.

2 Organisational aspects

Further aspects regarding data protection regulations, especially of an organisational nature including, for example, the questions of whether the data used are anonymised or pseudon-

(Bundesagentur für Arbeit - BA) this therefore applies only for research in the subject area of the labour market etc.

ymised, which players have to be involved and to what extent, as well as whether it is necessary to involve a trust agency.

2.1 Data access: anonymisation versus pseudonymisation

The legislator in Germany distinguishes between the anonymisation and pseudonymisation of data.

In this respect anonymisation is

> *the modifying of social data in such a way that the particulars about personal or factual circumstances can no longer be attributed to an identified or identifiable natural person or that this can only be done with a disproportionate amount of time, expense and effort. (§ 67 (8) SGB X) [see also § 3 (6) BDSG].*

The expression “no longer” refers here to the definition of absolute anonymity, as de-anonymisation is ruled out for certain here. The expression “with a disproportionate amount of time, expense and effort” refers to factual anonymity, which does not rule out de-anonymisation with absolute certainty. In the case of anonymized data, identification of the original person is no longer possible (absolute anonymity) or is only possible with a disproportionate amount of time and effort (factual anonymity). Data can be rendered anonymous, for example, by deleting details that can be attributed to the individual (e.g. age details) or by means of aggregation (e.g. providing age groups instead of date of birth or age, and only the month and year instead of the full date of a prescription). Case-related details are frequently regarded as anonymous.

Pseudonymisation is

> *if the name or another identifier is replaced by a substitute, so that the identification of the person is either impossible or at least rendered considerably more difficult (§ 67 (8a) SGB X) [see also § 3 (6a) BDSG].*

Pseudonymised data always fall under the sphere of data protection. In this case, certain personal identifiers (e.g. name, address, social security number) are deleted from the data or are replaced by “neutral” non-descriptive study identifiers (such as key identifiers), and “visible” characteristics are aggregated (such as dates of birth DD/MM/YYYY to years of birth YYYY), so that the researchers are given no so-called “unique information”. A key variable, which makes it possible to merge the personal identifiers with the pseudonymised data, may be deposited with a trust agency, if necessary (see Section 2.5).

The “advantage of pseudonymisation is that the pseudonym remains unchanged across the entire observation period, such that a patient’s data [from different data deliveries and] over longer observation periods can be merged to form patient-related histories and analysed” (Ihle 2005, p. 196).

Specific demands have to be met for the pseudonymisation procedure. First, the pseudonym must not contain any details that can be related to an individual or parts of such details

(Metschke und Wellbrock 2000). It is equally inadmissible to use a hash value, i.e. a pseudorandom number that is calculated directly from an original detail, e.g. the social security number. Although it is, by definition, technically virtually impossible to transform the hash value back into the original data, by forming the hash value of all the possible social security numbers, of which there is a finite number, it is possible to create a reference list of social security numbers and hash values, which can then be used to trace an existing hash value back to the social security number. In simple terms, a pseudonym should be formed in such a way that it is not possible to calculate or infer the original data from the pseudonym. There are various conceivable ways to do this, for example by compiling a reference list of original details and consecutive, non-descriptive numbers as pseudonyms. Alternatively, generating a pseudonym from the original data using an algorithm has proved effective. This has the advantage that there is no need for a reference list which then has to be kept secret. Only the key used must be stored in a safe and confidential place, so that even if an attacker knows the algorithm used, he will not be able to generate the pseudonym belonging to an existing social security number or to de-pseudonymise an existing pseudonym. The use of a reference list is a possible solution when it is only necessary to perform a pseudonymisation once; an algorithm, on the other hand, has advantages in cases of repeated pseudonymisation as the pseudonym can be calculated identically from the original value each time.

Before the pseudonymisation procedure it is often necessary to check the original data for homonyms and synonyms and to eliminate these. As is customary for epidemiological cancer registers, this function is also termed a trust agency. These two functions – comparison and pseudonymisation – can either be conducted by two institutions that are separate both in spatial and organisational terms, or they can be performed by one and the same institution; in the latter case, however, the two functions must be separated in terms of personnel and responsibility.

The secure and confidential storage of the reference list or the key is the actual task of a pseudonymisation agency. Confidential in the cryptographic sense means that the task of storage is entrusted to precisely one person or one agency that is solely responsible for it. For the scenarios described below, this has the consequence that a trust agency is to be set up as a pseudonymisation agency whenever at least two institutions supplying data are involved in the system. The actual pseudonymisation is purely a machine function and can be performed with few resources in terms of time, staff and funds using appropriate software in the sense of a pseudonymisation service (see Section 3.2).

There are different model scenarios for linking the data and pseudonymising personal details before the data are transmitted to the agency wishing to analyse them; these scenarios are presented in the following.

Scenario A: Pseudonymisation by the institution supplying the data

When data are supplied by one or more institutions, each with its own stock of insured persons, for example several statutory health insurance providers, then the data suppliers can pseudonymise their own data themselves. In this case, however, identical service or benefit providers will be given two separate pseudonyms if they are pseudonymised by two different

institutions. A central trust agency as a pseudonymisation body is not necessary here. The reference lists or keys used are stored by each data-supplying institution independently.

Scenario B: Central pseudonymisation

In this scenario, too, several data-supplying institutions are involved, but during the pseudonymisation process identical original details are to be given an identical pseudonym. This makes it essential to implement a central trust agency to conduct the pseudonymisation. In this scenario, for example, it is guaranteed that the social security number, which is allocated once and remains unchanged, is then also given the same pseudonym if the insured person changes to a different statutory health insurance provider.

In scenario B it may be necessary to compare the details that identify the individuals if several data sources supply data about one proband and different numbers or personal identifiable plain-text details are used to denote the proband. Showing these details on a unique study identifier that serves as a basis for the pseudonym requires a different amount of time and effort depending on the data situation and in some cases may require sophisticated linkage procedures. In cases of doubt it is also necessary to contact the data suppliers, which is resource-intensive.

Scenario C: Absolutely anonymised data basis

When making available a data basis that has been absolutely anonymised, for example by means of aggregation, the details that can be used to identify an individual have been eliminated entirely. As a result of the absolute anonymisation, the data basis no longer falls under the sphere of data protection regulations and is therefore unobjectionable as regards data protection legislation.

2.2 Actors

With regard to data protection, there are further actors that are of importance besides the researchers (of a research institution or a research consortium): the respective data protection officers (those of the data owners, those at national and regional level, and data protection officers belonging to the institutes), the data owners (e.g. health insurance providers), supervisory authorities responsible at national and regional level, ethics committees (e.g. at universities) as well as research data centres that provide access to data for researchers on a centralised basis.

Data protection officers are the people in an institution who are responsible for issues associated with data protection. They can be found both in the respective institutions that own the data, e.g. the health insurance providers, as well as at national or state level, for example in the form of the Federal Commissioner for Data Protection and the individual State Commissioners for Data Protection. Depending on the research project, it must be clarified in advance which data protection officers are to be involved in the planning or which jurisdiction the project comes under. If there is more than one data owner, it is also helpful to work to-

gether with all the data protection officers at an early stage, for despite the legal basis being the same, different views could be represented. Furthermore, it is necessary to involve the institute's own data protection officer, too. Good Practice in Secondary Data Analysis (Gute Praxis Sekundärdatenanalyse - GPS) continues to demand in Recommendation 8.10 that an additional person be named as being responsible for data protection besides the actual data protection officer of the particular institution, e.g. in the context of research projects (AGENS 2012).

Depending on the type of research project, it may be necessary to submit a request for permission to transmit data in accordance with § 75 SGB X to the relevant German supervisory authority (see Section 1.2). The German Federal Ministry of Labour and Social Affairs (Bundesministerium für Arbeit und Soziales - BMAS), for example, is the authority responsible for granting permission to transmit the data of the German Federal Employment Agency (BA); for statutory health insurance providers at national level it is the German Federal Insurance Authority (Bundesversicherungsamt - BVA), for statutory health insurance providers at federal state level it is the respective state supervisory authority. Information about data protection and about the responsible supervisory authority can be found in the legal notice on the respective individual data owner's website.

2.3 Research data centres

A number of research data centres (RDC) provide access to administrative data and/or survey data for researchers (Nimptsch et al. 2014). Most of the RDCs in Germany are accredited by the German Data Forum (Rat für Sozial- und Wirtschaftsdaten) if they meet certain minimum standards (Rat für Sozial- und Wirtschaftsdaten [no year], see also reference to the Research Data Centre (FDZ) of the BA at the Institute for Employment Research (IAB), Heining 2010).

The comparable access routes in all the RDCs differ with regard to the degree of anonymisation of the data that are made available:

- On-site use (access to pseudonymised or factually anonymous data taking a number of data protection regulations into account)
- Remote data access[3] (development of evaluation programs by researchers using test data, which are then run at the RDC using the available microdata)
- Scientific Use Files (factually anonymous datasets)
- Campus Files (absolutely anonymous datasets).

Legal regulations have to be observed depending on the type of access (see also Hochfellner et al. 2012). On-site access at the FDZ of the BA at the IAB is regulated in § 75 SGB X, and the transmission of Scientific Use Files in § 282 (7) SGB III. There are minimum re-

[3] In remote data access a distinction is made between remote access (where the researcher can also see the data) and remote execution (where the researcher has no direct contact with the data).

quirements that have to be met for each type of access: for on-site use, for example, the title of the specific research project has to be provided, as well as a description of the project including objectives, basic hypotheses and methods; the project's connection to the social security system has to be established, and the substantial public interest in the project and the necessity of the data and the particular variables/modules have to be justified. A project typically last no longer than five years. Before the transmission of Scientific Use Files, the person or institution submitting the request for data access also has to present a sophisticated data security concept that regulates the handling of personal data in the respective institutions (see next section). Furthermore, the output/results that are intended for publication must always fulfil the requirement of absolute anonymity; this is checked by the FDZ if the output/results were not created on the basis of Scientific Use Files.

2.4 Documents of relevance for data protection

Different documents pertaining to data protection are necessary depending on the research question, the data source type and the individuals involved (see also Section 1.2). These include, on the one hand, documents required for the respondents in a survey, such as information material, written declarations of consent ('informed consent') or a declaration of confidentiality. On the other hand, the researchers or groups of researchers involved in the project have to prepare various documents, for example an additional data protection agreement concluded by the cooperation partners, requests for the transmission of social data for a research project in accordance with § 75 SGB X, a data protection declaration, and data protection concepts of the individual parties involved.

The latter is also demanded by the GPS in recommendation 8.9:

> *All the data protection aspects necessary in the context of a secondary data analysis are to be set down in writing in a data protection concept. This includes, above all, the aspects of data transmission, data management, the duration of data storage and rights of access. Ideally, the data protection concept should be a component of the contract concluded with the data owner. In the case of studies with longer durations, the data protection concept is to be updated if necessary, with the involvement of the respective data protection officer responsible. (AGENS 2012, p. 8)*

If several institutions are involved in a research project, it would be appropriate to conclude an additional joint data protection agreement. This regulates all the consortium's data protection issues, including who is to be granted access to which data and for what period of time, when which data are to be deleted and how. In addition, the individual consortium partners' own data protection concepts are part of this agreement. Furthermore, for some institutions such an agreement is also a component of the request for the transmission of data in accordance with § 75 SGB X. Here, the respective institutions are required to submit a data security concept containing a description of the admission control, access control and data access control as well as dissemination control and job control at the institution submitting the request. In this context it is also important, first, that the data are used solely for the

named research project and, second, that the data are deleted fully after the end of the project. Retention of the data is not permitted.

The terms data protection concept and data security concept may possibly be used as synonyms. However, it is necessary to check whether the data protection concept also meets the content-related requirements of a data security concept as stipulated in the context of a request for the transmission of social data in accordance with § 75 SGB X should such a request be necessary for the particular research project. This should be checked as early as possible.

Furthermore, when routine data are transmitted, deletion deadlines, which are to be part of the respective data use agreements, must be observed. The Recommendation 8.6 of the GPS (as well as the principles of good scientific practice (Deutsche Forschungsgemeinschaft 2013); here: data kept for 10 years to ensure traceability) includes the storage of an unchanged original dataset and the analysis dataset and the observation of the corresponding retention period (AGENS 2012). The storage can, for example, be undertaken by a trust agency.

In addition, every person who works with personal data must be informed about the data protection legislation issues and must sign a data protection declaration as well as a declaration of confidentiality (Recommendation 8.11 of the GPS, AGENS 2012).

2.5 Trust agency

If the data of different data owners are to be linked in a research project, or if datasets are to be stored with personal identifiers, it is necessary to set up a trust agency (often called a data custodian). Besides the dissemination of pseudonymised (as well as anonymized) data, the tasks of this agency include above all the storage of the personal identifiers and the key variables that permit the linking of subdatasets. To this end, the trust agency must be shielded from all forms of external attacks, both in terms of personnel and in terms of IT technology; no unauthorised access (virtual or personal) may be permitted and absolute confidentiality must be given. The tasks and obligations of this agency must be laid down in a contract. For example, any access to the key files must be documented and the data protection officer responsible must be informed. Researchers are not allowed access. These demands are also reflected in the GPS, which in Recommendation 8.5 (AGENS 2012) demands special compliance with data protection regulations when personal data are linked with external data sources. In this case, too, a request for the transmission of data in accordance with § 75 SGB X must be submitted to the relevant supervisory authority providing detailed information as to which different data sources are to be linked and the research question for which the data are required.

Note: the term trust agency is utilised in different ways and is frequently used as the generic term for different functions, such as data acceptance, comparison of identifiers or pseudonymisation. If the agency deals solely with the function of pseudonymisation and the storage of reference lists, it may also be called a pseudonymisation agency. If an institution supplies pseudonymised data, then this pseudonymisation agency can also be located within that

institution and is represented by one specific person (plus their deputy) (see Section 2.1). The competencies and tasks of the trust agency are to be defined in project-specific or institution-specific terms. For further information see Ihle (2005, 2008), March et al. (2012) and AGENS (2012).

3 Two examples of studies taking account of data protection legislation

3.1 lidA – leben in der Arbeit – study

lidA – leben in der Arbeit[4], is a German cohort study on health and ageing in working life and examines the long-term (current and future) effect of work on the health of an ageing labour force in Germany. The study is conducted by a consortium (the universities of Wuppertal, Magdeburg and Ulm, the Institute for Employment Research and infas Institute for Applied Social Sciences as well as the Federal Institute for Occupational Safety and Health as an associated partner; www.lida-studie.de). The heart of the study is a survey (comprising at least two waves) of 6,585 employees belonging to the birth cohorts of 1959 and 1965. The cohorts are representative of the German labour force of the same age (excluding the self-employed and civil servants; Schröder et al. 2013). The survey data are merged with administrative data from the BA / the IAB (the so-called Integrated Employment Biographies/IEB) (Oberschachtsiek et al. 2009). The IEB contain the full employment biographies, on a day-to-day basis, of the people in employment covered by social security in Germany (no information on civil servants and the self-employed). The sample for the survey was also drawn from these data. On the other hand, these data are linked with individual health insurance data (claims data) from various sectors of the statutory health insurance providers (as well as aggregate health insurance data of the two cohorts in the context of a work-health matrix) (see March et al. 2014).

The linking of the respective administrative data (IEB and statutory health insurance data) and the inclusion in the panel (repeat survey) were only performed, however, if the respondents had provided written consent for each individual step (Schröder et al. 2013). When asking the respondents for their consent it was necessary to take into account that this constitutes “informed consent”. In order to achieve this, the respondents were given detailed participant information regarding various aspects, including the aims of the project, descriptions of the data to be linked, a data protection concept, as well as the method of address storage (for the follow-up survey). The respondents were also explicitly informed that they could revoke their consent at any time (ibid.).

In order to comply with the data protection regulations, requests for permission to transmit data in accordance with § 75 SGB X were submitted to the relevant supervisory authorities in the run-up to the study: a request submitted to the BMAS for permission to draw the sample and to transmit the address data (data basis: IEB), a request submitted to the German Federal Ministry of Health (Bundesministerium für Gesundheit – BMG, responsible until 31.12.2011) or to the German Federal Office of Administration (Bundesverwaltungsamt –

[4] The study is funded by the Federal Ministry of Education and Research (BMBF) (funding codes of the projects involved in the consortium: 01ER0806, 01ER0825, 01ER0826, 01ER0827).

BVA, responsible from 01.01.2012 onwards) in the case of statutory health insurance providers acting throughout Germany, and one to the federal state authorities in the case of regional statutory health insurance providers, for permission to make available and transmit the statutory health insurance data. The cooperation with several health insurance providers made it necessary to coordinate and include all the individuals involved, e.g. the individual data protection officers, fully at an early stage. Although each health insurance provider, as a data owner, has to submit its own request for permission to transmit data in accordance with § 75 SGB X to their supervisory authority, it is helpful in this context, for example, to provide all the health insurance providers involved with a sample request and to provide active assistance.

In addition to this, the cooperation partners have concluded a comprehensive additional data protection agreement, which regulates, among other things, data transmission and data management (the data security and data protection concepts of each institution). The linkage of the data stocks is performed by a trust agency which stores the key identifiers in a separate system and provides the researchers only with pseudonymised data. Every case of access to the key files must be documented and the data protection officer must be informed (March et al. 2012).

3.2 Statutory Health Insurance Sample of Hesse

The regional Statutory Health Insurance Sample of Hesse[5] (Versichertenstichprobe AOK Hessen/KV Hessen) (Ihle et al. 2005) is a random sample of people insured with the AOK statutory health insurance fund in Hesse, which has been drawn continuously since the survey year 1998 (currently extending to 2012). As the data are drawn separately from the data stocks of the AOK and from the Association of Statutory Health Insurance Physicians in Hesse (Kassenärtzliche Vereiningung – KV) and are linked at pseudonymised level, a specific data protection concept was agreed with the data protection officer of the federal state of Hesse and the data protection officers of the participating institutions. The pseudonymisation service, which is located at a separate trust agency, was implemented as a cost-effective routine solution. The PMV research group (University Hospital of Cologne) developed the necessary software for this with financial support from the TMF – Technology, Methods and Infrastructure for Networked Medical Research (TMF e. V. Berlin) (Ihle 2004). The pseudonymisation is scalable and can therefore be adapted to different dataset structures. In addition to characteristics related to the insured persons, such as their health insurance number, it is also possible to pseudonymise institution-related details such as physicians' registration IDs or institutional identifiers of in-patient establishments or case numbers derived from these for the purpose of the study. The scaling parameters are defined by the institution supplying the data and, together with the data that are to be pseudonymised, they are transmitted to the trust agency, which in this case serves as a pseudonymisation body. In this connection the medical data are passed though in encrypted form; the pseudonymisation service sees only the top of the data records, which contains the identification data that are to be pseudonymised. The data are only stored temporarily for the duration of the pseudonymisation pro-

[5] Hesse (Hessen) is one of the German federal states (Bundesländer)

cess. The transmission of the data to the pseudonymisation agency and the subsequent transmission to the evaluation body (here: the researcher) following the pseudonymisation process is done offline and using transport encryption (Ihle 2005).

> *The pseudonymisation service for routine data has made it possible to implement a tested procedure that is unobjectionable as regards data protection both in organisational and technical terms. In routine use, the procedure has proved to be stable, practical and inexpensive. The service can also be used in other environments, e.g. in export and for converting clinical databases into scientific databases. (ibid. 2005, p. 200)*

4 Conclusion

When using German social data or routine data in scientific research it is necessary to comply with extensive data protection legislation. In order to be able to use these data – and this is a statement, which is true in an international perspective –, the amount of organisation and time required must already be taken into account in the planning phase in the context of cost-benefit considerations. This applies not only for the use of individual data sources but also for the linking of different data sources (e.g. primary and secondary data).

Nonetheless, despite all the organisational and legal challenges illustrated in this paper, it is still worth the effort. Social data, both used alone as well as in combination with survey data, will continue to gain in importance as a basis for future research projects. It is important in this respect, however, that the data protection standards are observed and continue to be developed further in the light of the changing demands and basic conditions. Examples of how these data may be used are provided in various chapters in Swart et al. (2014). Many of the aspects presented here can also be found in the GPS (AGENS 2012).

References

AGENS (2012): Gute Praxis Sekundärdatenanalyse (GPS) Leitlinien und Empfehlungen. 3. Fassung http://dgepi.de/fileadmin/pdf/leitlinien/GPS_fassung3.pdf

Bundesbeauftragter für den Datenschutz und die Informationsfreiheit (2013): Tätigkeitsbericht zum Datenschutz für die Jahre 2011 und 2012 http://www.bfdi.bund.de/SharedDocs/Publikationen/Taetigkeitsberichte/TB_BfDI/24TB_11_12.pdf?__blob=publicationFile

Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert. http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf

Deutsche Forschungsgemeinschaft (2013): Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Proposals for safeguarding good scientific practice. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

Grundgesetz für die Bundesrepublik Deutschland vom 23. Mai 1949 (BGBI. I S. 1) in der im Bundesgesetzblatt Teil III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, zuletzt durch Artikel 1 des Gesetzes vom 11. Juli 2012 (BGBl. I S. 1478) geändert. http://www.gesetze-im-internet.de/bundesrecht/gg/gesamt.pdf

Heining J (2010): The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009. Zeitschrift für ArbeitsmarktForschung, 42/4: 337–350.

Hochfellner D, Müller D, Schmucker A, Roß E (2012): Datenschutz am Forschungsdatenzentrum. FDZ-Methodenreport 06/2012. IAB, Nürnberg.

Ihle P (2004): Ergebnisbericht und Manual. Pseudonymisierungsdienst. „Sekundärdaten“ Implementierung eines Pseudonymisierungsdienstes mit Treuhänderstelle und Erstellung einer Pseudonymisierungssoftware unter besonderer Berücksichtigung der Anforderungen bei der Pseudonymisierung von Gesundheits- und Sozialdaten für die Sekundärdatenanalyse – Darstellung der Organisationsstruktur, Programmbeschreibung und Installationsanleitung. Teilprojekt im Projekt DS 3.1 „Pseudonymisierungsdienst“ der Arbeitsgruppe „Datenschutz und Datensicherheit“ der Telematikplattform für medizinische Forschungsnetze (TMF), Version 1.01.

Ihle P (2005): Datenschutzrechtliche Aspekte bei der Erhebung von GKV-Routinedaten. In: Swart E, Ihle P (Hrsg.): Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. Verlag Hans Huber, Bern: 195–201.

Ihle P (2008): Datenschutzrechtliche und methodische Aspekte beim Aufbau einer Routinedatenbasis aus der Gesetzlichen Krankenversicherung zu Forschungszwecken. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 1: 1127–1134.

Ihle P, Köster I, Herholz H, Rambow-Bertram P, Schardt T, Schubert I (2005): Versichertenstichprobe AOK Hessen/KV Hessen – Konzeption und Umsetzung einer personenbezogenen Datenbasis aus der Gesetzlichen Krankenversicherung. Gesundheitswesen 67: 638–645.

John J, Krauth C (2005): Verknüpfung von Primärdaten mit Daten der Gesetzlichen Krankenversicherung in gesundheitsökonomischen Evaluationsstudien. Erfahrungen aus zwei KORA-Studien. In: Swart E, Ihle P (Hrsg.): Routinedaten im Gesundheitswesen. Handbuch

Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. Verlag Hans Huber, Bern: 215–234.

Lüdtke O, Tomasik MJ, Lang FR (2003): Teilnahmewahrscheinlichkeit und Stichprobenselektivität in altersvergleichenden Erhebungen. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 35 (3): 171–180.

March S, Rauch A, Thomas D, Bender S, Swart E (2012): Datenschutzrechtliche Vorgehensweise bei der Verknüpfung von Primär- und Sekundärdaten in einer Kohortenstudie: die lidA-Studie. Gesundheitswesen 74 (12): 834–835 [Langfassung: Gesundheitswesen 74 (12): e122–e129].

March S, Stallmann C, Swart E (2014): Datenlinkage. In: Swart E, Ihle P, Gothe H, Matusiewicz D (Hrsg.): Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 2., vollst. überarb. Aufl. Verlag Hans Huber, Bern: 347–355.

Metschke R, Wellbrock R (2000) Datenschutz in Wissenschaft und Forschung. Materialien zum Datenschutz Nr. 28 Berliner Beauftragter für den Datenschutz und Akteneinsicht. 2. überarbeitete Auflage. Verwaltungsdruckerei, Berlin.

Nimptsch U, Bestmann A, Erhart M, Dudey S, Marx Y, Saam J et al. (2014): Zugang zu Routinedaten. In: Swart E, Ihle P, Gothe H, Matusiewicz D (Hrsg.): Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 2., vollst. überarb. Aufl. Verlag Hans Huber, Bern: 270–290.

Oberschachtsiek D, Scioch P, Seysen C, Heining J (2009): Stichprobe der Integrierten Erwerbsbiografien IEBS. Handbuch für die IEBS in der Fassung 2008. FDZ-Datenreport, 03/2009.

Rat für Sozial und Wirtschaftsdaten (ohne Jahr): Akkreditierte Datenzentren – ein vielfältiges Datenangebot. http://ratswd.de/forschungsdaten/fdz

Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr. Amtsblatt Nr. L 281 vom 23/11/1995 S. 0031–0050. http://eur-lex.europa.eu/legal-content/de/ALL/?uri=CELEX:31995L0046

Schnell R (1997): Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Leske + Budrich Verlag, Opladen.

Schröder H, Kersting A, Gilberg R, Steinwede J (2013): Methodenbericht zur Haupterhebung lidA – leben in der Arbeit. FDZ-Methodenreport 01/2013.

Sozialgesetzbuch Erstes Buch (SGB I) – Allgemeiner Teil – (Artikel I des Gesetzes vom 11. Dezember 1975, BGBl. I S. 3015), zuletzt durch Artikel 10 des Gesetzes vom 19. Oktober 2013 (BGBl. I S. 3836) geändert. http://www.gesetze-im-internet.de/bundesrecht/sgb_1/gesamt.pdf

Sozialgesetzbuch Drittes Buch (SGB III) Arbeitsförderung – (Artikel 1 des Gesetzes vom 24. März 1997, BGBl. I S. 594, 595), zuletzt durch Artikel 11 des Gesetzes vom 19. Oktober 2013 (BGBl. I S. 3836 geändert. http://www.gesetze-im-internet.de/bundesrecht/sgb_3/gesamt.pdf

Sozialgesetzbuch Fünftes Buch (SGB V) – Gesetzliche Krankenversicherung – (Artikel 1 des Gesetzes v. 20. Dezember 1988, BGBl. I S. 2477, 2482), durch Artikel 3 des Gesetzes vom

7. August 2013 (BGBl. I S. 3108) geändert. http://www.gesetze-im-internet.de/bundesrecht/sgb_5/gesamt.pdf

Sozialgesetzbuch Zehntes Buch (SGB X) – Sozialverwaltungsverfahren und Sozialdatenschutz – in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), zuletzt durch Artikel 6 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749) geändert. http://www.gesetze-im-internet.de/bundesrecht/sgb_10/gesamt.pdf

Strafgesetzbuch in der Fassung der Bekanntmachung vom 13. November 1998 (BGBl. I S. 3322), zuletzt durch Artikel 5 Absatz 18 des Gesetzes vom 10. Oktober 2013 (BGBl. I S. 3799) geändert. http://www.gesetze-im-internet.de/bundesrecht/stgb/gesamt.pdf

Swart E, Ihle P, Gothe H, Matusiewicz D (Hrsg.) (2014): Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven. 2., vollst. überarb. Aufl. Verlag Hans Huber, Bern.

von Ferber C (1997): Die Herausforderungen der Public Health Forschung durch den Datenschutz. In: von Ferber L, Behrens J (Hrsg.): Public Health Forschung mit Gesundheits- und Sozialdaten. Stand und Perspektiven. Asgard-Verlag, Sankt Augustin: 193–208.

Imprint

FDZ-Methodenreport 12/2015

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dr. Jörg Heining, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2015/MR_12-15_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Stefanie March
Institut für Sozialmedizin und
Gesundheitsökonomie (ISMG)
Med. Fakultät
Otto-von-Guericke-Universität Magdeburg
Leipziger Str. 44
39120 Magdeburg
email: stefanie.march@med.ovgu.de