Published online: 2019-12-02

Linkage of Routine Data to Other Data Sources in Germany: A Practical Example Illustrating Challenges and Solutions

Verknüpfung von Routinedaten mit anderen Datenquellen in Deutschland: Herausforderungen und Lösungen dargestellt an einem praktischen Beispiel

Authors
Ingo Langner[1], Oliver Riedel[1], Jonas Czwikla[2, 3], Franziska Heinze[2, 3], Heinz Rothgang[2, 3], Hajo Zeeb[3, 4], Ulrike Haug[1, 3]

Affiliations
1 Klinische Epidemiologie, Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS, Bremen
2 SOCIUM Forschungszentrum Ungleichheit und Sozialpolitik, Universität Bremen, Bremen
3 Wissenschaftsschwerpunkt Gesundheitswissenschaften, Universität Bremen, Bremen
4 Abt. Prävention und Evaluation, Leibniz-Institut für Präventionsforschung und Epidemiologie, Bremen

Key words
secondary data, claims data, record linkage, cause of death, routine data

Schlüsselwörter
Sekundärdaten, Krankenkassendaten, Datenabgleich, Todesursache, GKV-Routinedaten

Bibliography
DOI https://doi.org/10.1055/a-0999-5509
Online-Publikation: 2.12.2019
Gesundheitswesen 2020; 82 (Suppl. 2): S117–S121

ISSN 0949-7013

Correspondence
Dr. Ingo Langner
Klinische Epidemiologie
Leibniz-Institut für Präventionsforschung und Epidemiologie – BIPS,
Achterstraße 30
28359 Bremen
Germany
langner@leibniz-bips.de

ZUSAMMENFASSUNG

Routinedaten haben ein hohes Potenzial für die epidemiologische und Versorgungsforschung, doch fehlen beispielsweise Informationen zu Todesursachen und es mangelt häufig an detaillierten Informationen u. a. zu Lebensstilfaktoren. In Deutschland ist v. a. aufgrund der strengen Datenschutzauflagen eine unmittelbare Ergänzung dieser Informationen durch andere Datenquellen („Linkage") mit einigen Herausforderungen verbunden. So besteht eine Herausforderung darin, dass die Routinedatennutzer üblicherweise keinen Zugriff auf Personenidentifikatoren besitzen, die für ein Linkage erforderlich sind. Darüber hinaus sollten sensible Informationen (z. B. die Todesursache) nicht an Institutionen übermittelt werden, die selbst über Personenidentifikatoren verfügen. In diesem Artikel veranschaulichen wir diese zentralen Herausforderungen und stellen entsprechende Lösungen anhand eines Praxisbeispiels vor, bei dem die Abrechnungsdaten von gesetzlichen Krankenkassen mit einem epidemiologischen Krebsregister verknüpft werden, um Informationen zur Todesursache zu erhalten. Wir beschreiben die für das Linkage notwendigen Genehmigungsverfahren, den Datenfluss zwischen den beteiligten Institutionen und erläutern die Gründe für den Datenfluss im Hinblick auf die wesentlichen Herausforderungen. Schließlich verallgemeinern wir die Fragen, die bei der Planung eines Linkage-Verfahrens zu klären sind und zeigen mögliche weitere Herausforderungen auf. Mit diesem praktischen Beispiel zeigen wir, dass ein Abgleich zwischen Routinedaten und anderen Datenquellen in Deutschland umsetzbar ist, dabei aber bestimmte Beschränkungen und Hindernisse berücksichtigt werden müssen.

ABSTRACT

Routine data have a high potential for epidemiological and health care research but lack information, for instance, on the cause of death. Often detailed information, such as on lifestyle factors is also missing. In Germany, obtaining the missing information by linkage to data sources is challenging, mainly due

to strict data protection regulations. One key challenge arises from the fact that routine data users usually have no access to person identifiers which would be necessary for record linkage. A second key challenge is that sensitive information (i. e., the cause of death) should not be transferred to an institution that holds person identifiers. In this paper, we illustrate these key challenges and present corresponding solutions based on a practical example where claims data from statutory health insurance providers are linked to an epidemiological cancer registry to obtain cause of death information. We describe the approval procedures necessary for the record linkage, the dataflow between the involved institutions and explain the rationale of the dataflow in view of the key challenges. Finally, we generalize the questions that need to be addressed when a record linkage is planned and point to additional potential challenges. Overall, we illustrate that a linkage between routine data and other data sources in Germany is feasible, but specific restrictions and hurdles need to be taken into consideration.

Background and Challenges

Routine data are of increasing importance in epidemiological and healthcare research [1, 2]. They provide population-based information on a large number of persons, avoid non-response and recall bias, and often allow for long follow-up periods with a constant degree of detail. However, routine data may also lack information that is important for certain research questions, such as detailed information on lifestyle or sociodemographic factors, genetic and other molecular information and the cause of death (CoD).

One way to overcome this limitation of routine data is the linkage to data sources containing the missing information [3], which is in principle feasible if the data are not anonymized. For such a linkage, both data sources have to include certain identifiers with which records pertaining to the same person can be linked. However, there are several challenges to such a data linkage in Germany, mainly resulting from strict data protection regulations [4, 5]. Individual level routine data are sensitive and their storage, access and use is therefore tightly regulated by law. As a consequence, before data holders (i. e., Statutory Health Insurance funds (SHIs)) provide routine data to be used for research purposes they must at least pseudonymize the data because routine data users (RDUs) are not allowed to have any person identifiers (i. e., name, address). Given that most data holders have their own confidential pseudonymization method the pseudonyms vary between different data sources and the RDU can thus not conduct a direct record linkage between these data sources. Accordingly, a solution may be that the record linkage is conducted by the data provider where person identifiers are available but there are also hurdles concerning the "deepness" of the data. It would, for example, not be allowed to hold – at the individual level – person identifiers together with claims data and information on the CoD at the same institution.

Thus, one key challenge of linking routine data to other data sources is to establish a dataflow that incorporates the person identifiers necessary for linkage while making sure that the RDU will not gain access to these identifiers. A second key challenge in this dataflow is to avoid the transfer of sensitive information (i. e., the cause of death) to an institution that holds person identifiers.

In this paper, we illustrate these key challenges and present corresponding solutions based on a practical example where claims data from SHIs are linked to an epidemiological cancer registry (ECR) to obtain CoD information. This is relevant in the context of a study that will evaluate the German Mammography Screening Program based on claims data, requiring information on breast cancer deaths [6–8]. Following this practical example, we will generalize the potential solutions to other data sources.

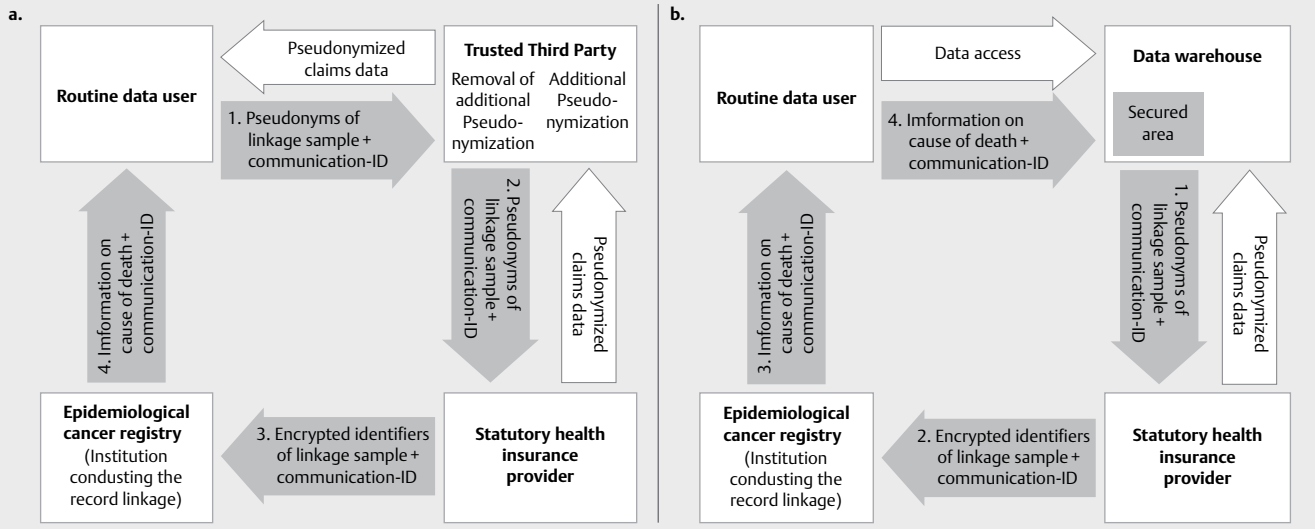
Practical Example: Linkage of Claims Data with an Epidemiological Cancer Registry

The aim is to link claims data from different SHIs to the ECR of North Rhine Westphalia (NRW) in order to obtain information on the CoD. Three institutions are involved in the linkage: the RDU, the SHI holding person identifiers for the claims data and the ECR as exporter of information on the CoD and linking institution. With respect to the SHIs, there are two different ways of providing the RDU access to the data as illustrated by the white arrows in ▶ **Fig. 1a** and **b:** A) pseudonymized data are transferred from the SHI to the RDU via a third trusted party (which applies a second pseudonymization); B) there is access to pseudonymized data in a data warehouse of the respective SHI via virtual connection, i. e., the data are not transferred to the RDU. Given that the kind of data access determines the dataflow, both scenarios will be explained. In the following, we will first describe the approval procedures necessary for the record linkage, then describe the dataflow between the involved institutions and finally explain the rationale of the dataflow in view of the key challenges.

Approval procedures

In a first step, the access to claims data itself by the RDU requires approval – irrespective of the additional approval required for the data linkage. If the data is not transferred to the RDU – as is the case with data access B – it is sufficient to have the approval of the respective SHI to access the data. If the data is transferred to the RDU – as is the case with data access A, the approval of the respective SHI is a prerequisite, but not sufficient. In Germany, the transfer of claims data for research purposes is regulated by Article 75, Book 10 of the Code of Social Law (Sozialgesetzbuch, SGB X), saying that the governing authority of the SHI has to grant approval for transferring claims data to the RDU to be used for a specific research question. For nationwide operating SHIs, the governing authority is the German Federal Insurance Office.

In a second step, the data linkage itself requires approval by the governing authority. This holds true for both ways of data access (A and B) given that transfer of data is necessary for the linkage. Accordingly, a project proposal and a data protection concept which includes a detailed description of the dataflow (see below) were sent to the governing authority of the SHIs involved in the data linkage.



▶ **Fig. 1** Dataflow between institutions, first from the statutory health insurance provider to the routine data user (white arrows) and subsequently for record linkage of claims data with cause of death information (grey arrows); **a** data access model A, **b** data access model B.

In addition, it requires approval that the ECR participates in the record linkage and provides information on the CoD. Approval procedures concerning research projects with the data of the ECR are regulated on the federal state level by the respective cancer registry act. In our example, the former Cancer Registry Act of NRW (valid until March 31st, 2016) permits the ECR NRW itself to approve such research projects and export CoD information for research purposes (Article 10), while in other federal states only the governing authority of the cancer registry is allowed to give approval. Accordingly, a corresponding request for the data linkage was sent to the ECR NRW directly.

Description of the dataflow for the record linkage

As illustrated in ▶ **Fig. 1a** (grey arrows) for data access A, in the first step of the dataflow the RDU selects the linkage sample—in this example deceased women 25–80 years old at the age of death with NRW residence—and transfers only the pseudonyms of the selected individuals together with consecutive numbering (communication-ID) via the trusted third party back to the SHI. The communication-ID is used as a quasi-anonymous identifier which can be passed through all institutions involved in the dataflow to enable a later addition of the required information (CoD in this example) to the RDU dataset. In the second step, the SHI adds the person identifiers needed for the record linkage to the sample. To guarantee that the person identifiers available at the SHI and the ECR could be linked, IT personnel of both institutions exchanged information to harmonize type, number, format, and quality of these identifying variables. As the ECR is allowed to handle double encrypted identifiers only [9, 10], a procedure is chosen which encrypts the SHI person identifiers corresponding to the requirements of the ECR. The ECR provides the SHI with a data-transfer and encryption software tool, which the SHI employ for the first encryption (MD5 method) of the identifiers and an additional transport encryption [11]. These encrypted data are transferred from the SHIs to the ECR via a specific pseudonymization service (implemented at the Kassenärztliche Vereinigung Westfalen-Lippe) which removes the transport encryption and applies the second encryption (IDEA method). The data transferred from the SHI to the ECR include identifiers only and no medical or other information. In the third step, the ECR conducts the record linkage based on the encrypted person identifiers using a probabilistic method [9], adds the CoD information for those individuals who could be matched, deletes all identifiers except for the communication-ID and sends the data to the RDU. In the fourth step, the CoD information is linked to the sample (selected in the first step) at the RDU via the communication-ID.

Regarding data access B, the dataflow is similar with two exceptions (▶ **Fig. 1b**). In the first step, the RDU conducts the sample selection and transfer of the pseudonyms via the data warehouse. For the fourth step, the RDU transfers the received CoD information to a secured area of the data warehouse (which is only accessible for the RDU) before linking the information to the sample.

Rationale of the dataflow in view of the key challenges

One key challenge is to establish a dataflow that incorporates the person identifiers necessary for the linkage while making sure that the RDU will not gain access to these identifiers. In our example, this is guaranteed by a combination of the following features of the dataflow: 1) the data transfer between institutions is only in one direction; 2) the SHI as the institution holding person identifiers for the claims data is involved in the data flow directly after the RDU only; 3) only institutions that are in possession of person identifiers anyhow were involved in the actual record linkage; 4) the (encrypted) person identifiers are deleted immediately after the record linkage, i. e. before the RDU is again involved in the data flow.

A second key challenge is to avoid the transfer of sensitive information (i. e., the cause of death) to an institution that holds person identifiers. In our example, the RDU starts the dataflow with a list of pseudonyms only, i. e., without any sensitive information on the

▶ **Table 1** Overview of key questions that need to be addressed and clarified when a linkage between routine data and other data sources is planned.

– Which institution is holding the information to be added to the routine data?
– Does this institution have permission to export the respective information? Which approval processes will have to be initiated for the export?
– Does this institution have permission to conduct a record linkage?
– Are the institutions involved in the data linkage able to provide (encrypted) person identifiers in a format that guarantees a suitable quality of the linkage? Which format is required?
– How does the dataflow need to be conceived to ensure that no institution receives information that is not allowed to be held by that institution (i. e., because this institution also holds person identifiers)?

selected individuals. The data flow from the SHI to the ECR is restricted to the (encrypted) identifiers required for the linkage and after the record linkage, the ECR transfers CoD information to the RDU only after the (encrypted) person identifiers have been removed. Thus, no institution holding person identifiers receive any sensitive information on the selected individuals from another institution in this data flow.

The further data handling differs between data access A and B. Regarding data access A, the RDU can directly link the CoD information to the claims data via the communication-ID. Regarding data access B, the RDU needs to transfer the CoD information to the data warehouse first. Since the SHI holding the person identifiers usually also has access to the data warehouse, a secured area within the data warehouse needs to be installed which is only accessible for the RDU. Storing the CoD information, linking the claims data with the CoD, as well as storing and processing all further data files that include the CoD information have to take place in this secured area.

Application to Other Data Sources and Potential Additional Challenges

After having illustrated challenges and solutions for a linkage between routine data and other data sources in Germany based on a practical example it is important to note that there is no standard solution that is applicable to all data sources.

When a record linkage is planned, a number of questions need to be systematically addressed and clarified (▶ **Table 1**). The first question is which institution holds the information that should be added to the routine data. Then it needs to be clarified whether this institution is allowed to conduct a record linkage and to transfer the information at an individual level to other data sources. There might be legal restrictions regarding such a data transfer. For example, the Federal Statistical Office in Germany holds information on the CoD at an individual level but does not have permission for a record linkage. In our example, the ECR NRW provided information on the CoD. Generally, cancer registries receive information on the CoD at an individual level to determine cancer deaths but only in some federal states (including NRW) the respective legislation allows the cancer registry to transfer this information to other institutions. Apart from that, some cancer registries are not allowed to keep CoD information from persons without cancer, but only for cancer patients recorded in the registry. Thus, it is not only relevant to identify an institution holding the required information but also to clarify whether the applicable (federal) law will allow the institution to export this information.

Once these basic issues are clarified, another question is whether the institutions providing person identifiers for the actual data linkage could do this in a format that guarantees a suitable quality of the linkage and if so, which format is best suited [3]. This question is less relevant in settings where each person has a unique numerical person identifier that is recorded in the various data sources such as in Scandinavian countries. In Germany, however, a linkage between different data sources typically needs to be done based on non-numerical person identifiers (name, address etc.) bearing the risk of mismatches. As in our example, it may be the case that one of the institutions involved in the linkage (in our example the ECR) is only allowed to handle encrypted data. While we described one solution to meet this requirement, there could also be alternative approaches [12, 13].

In our example, the record linkage was conducted by the institution holding the information to be added to the routine data. Technically, the linkage could also take place at the institution holding the person identifiers of the routine data (i. e., the SHI in our example). This approach, however, would mean that the transfer of (encrypted) identifiers between both institutions cannot be restricted to a linkage sample that has been selected based on information available in routine data beforehand. In our example, this would also have entailed that data on persons not insured at the participating SHI had been transferred to the participating SHI because the ECR has no information who is insured at which SHI. It seems doubtful whether such a data flow would be approved by the respective authorities.

Concluding Remarks

Our paper illustrates - based on a practical example - that a linkage between routine data and other data sources in Germany can be feasible, but specific restrictions and hurdles need to be considered. In general, the ways are not paved and require certainly more efforts than in several other countries. The fact that the RDU in Germany does typically not have access to person identifiers complicates the dataflow. Furthermore, this implies that the support and commitment of the institution initially providing the routine data (and holding person identifiers) is an essential prerequisite, in addition to the support required by the institution providing the lacking information. There might be situations where a data linkage does not receive enough priority at the respective institutions i. e., due to competing tasks and limited resources. To realize the high potential of linked routine data for epidemiological and healthcare research it will thus be important to have funding sources that allow to explore new approaches and to optimize existing options.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Hoffmann F. Review on use of German health insurance medication claims data for epidemiological research. Pharmacoepidemiol Drug Saf 2009; 18: 349–356

[2] Kreis K, Neubauer S, Klora M et al. Status and perspectives of claims data analyses in Germany-A systematic review. Health Policy 2016; 120: 213–226

[3] Hoffmann F, Abbas S. Gut gelinkt ist halb gewonnen: Es könnte alles so einfach sein, ist es aber nicht. Gesundheitswesen 2015; 77: 72–73

[4] March S, Antoni M, Kieschke J et al. [Quo Vadis Data Linkage in Germany? An Initial Inventory]. Gesundheitswesen. 2018; 80: e20–e31

[5] March S, Rauch A, Bender S et al. Data protection aspects concerning the use of social or routine data. FDZ Methodenreport 2015; 12: 1–22

[6] Fuhs A, Bartholomäus S, Heidinger O et al. [Evaluation of the impact of the mammography screening program on breast cancer mortality: feasibility study on linking several data sources in North Rhine-Westphalia]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57: 60e7

[7] Hense HW, Barlag H, Bartholomäus S et al. Evaluation der Brustkrebsmortalität im Deutschen Mammographie-Screening-Programm - Vorhaben. 3610S40002

[8] Czwikla J, Giersiepen K, Langner I et al. A cohort study of mammography screening finds that comorbidity measures are insufficient for controlling selection bias. J Clin Epidemiol 2018; 104: 1–7

[9] Giersiepen K, Bachteler T, Gramlich T et al. [Performance of record linkage for cancer registry data linked with mammography screening data]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2010; 53: 740–747. doi:10.1007/s00103-010-1084-1

[10] Meyer M. Kontrollnummern und Record-Linkage. In Hentschel S, Katalinic A, Hrsg. Das Manual der epidemiologischen Krebsregistrierung. München: Zuckschwerdt; 2008: 57–68

[11] Langner I, Krieg V, Heidinger O et al. Enrichment of claims data with official causes of death using a record linkage with the epidemiological cancer registry of north rhine-westphalia: feasibility study and comparison of procedures. Gesundheitswesen 2018, doi:10.1055/s-0043-124669. Epub ahead of print

[12] Ohlmeier C, Hoffmann F, Giersiepen K et al. Linkage of statutory health insurance data with those of a hospital information system: Feasible, but also “Useful”?. Gesundheitswesen 2015; 77: e8–e14. doi:10.1055/s-0034-1395644

[13] Maier B, Wagner K, Behrens S et al. Deterministic record linkage with indirect identifiers: Data of the Berlin Myocardial Infarction Registry and the AOK Nordost for patients with myocardial infarction. Gesundheitswesen 2015; 77: e15–e19. doi:10.1055/s-0034-1395642