Published online: 2020-03-19

Individual-Level Linkage of Primary and Secondary Data from Three Sources for Comprehensive Analyses of Low Birthweight Effects

Individuelles Datenlinkage von Primär- und Sekundärdaten aus drei Datenquellen zur umfassenden Analyse der Effekte eines geringen Geburtsgewichtes von Kindern

Authors
Diana Druschke[1‡], Katrin Arnold[1‡], Luise Heinrich[1], Jörg Reichert[2], Mario Rüdiger[2], Jochen Schmitt[3]

Affiliations
1 Zentrum für Evidenzbasierte Gesundheitsversorgung, Universitätsklinikum Carl Gustav Carus, Dresden, Germany
2 Klinik und Poliklinik für Kinder- und Jugendmedizin, Fachbereich Neonatologie und Pädiatrische Intensivmedizin, Universitätsklinikum Carl Gustav Carus, Dresden, Germany
3 Zentrum für Evidenzbasierte Gesundheitsversorgung, Universitätsklinikum Carl Gustav Carus an der Technischen Universität Dresden, Dresden, Germany

Key words
data linkage, claims data, secondary data, cohort study, preterm infants, data protection

Schlüsselwörter
Datenlinkage, Routinedaten, Sekundärdaten, Kohortenstudie, Frühgeborene, Datenschutz

Bibliography
DOI https://doi.org/10.1055/a-1082-0740
Gesundheitswesen 2020; 82 (Suppl. 2): S108–S116

ISSN 0949-7013

Correspondence
Diana Druschke
Zentrum für Evidenzbasierte Gesundheitsversorgung
Universitätsklinikum Carl Gustav Carus
Fetscherstr. 74
01307 Dresden
Germany
diana.druschke@uniklinikum-dresden.de

ABSTRACT

Aim of the study The linkage of primary and secondary data is becoming an increasingly popular approach in healthcare research, but involves some challenges for all involved parties, for example due to data protection requirements. The aim of this article is to systematically outline the methods used and experiences made during a cohort study in the field of pediatric health care research (EcoCare-PIn) that involved access to and linkage of three different data sources. Particular focus is placed on the necessary regulatory measures with regard to data access and data linkage as well as on data validation to ensure a correct linkage.

Methods While complying with all relevant data protection requirements, the study realized an individual-level linkage of a) pseudonymized administrative health insurance data from a statutory health insurance on Saxon children born between 2007 and 2013, b) primary data collected via postal questionnaires from parents/caregivers and c) medical data from kindergarten- and school-entry-examinations of Saxon health authorities. The fundamental principle of the concept of data linkage was to strictly separate the sites of data collection and data analysis, which was realized through the involvement of a trust center.

Results Challenges especially pertained to the extensive regulatory pre-requirements for data access as well as to data protection requirements while performing the study. Technical aspects and data validation also required a considerable share of attention and resources. A number of validation routines were applied to avoid incorrect data linkage and to ensure the high quality of the final dataset. Data validation included both plausibility checks within the primary data and consistency checks of information given in primary and secondary data.

Conclusion The linkage of primary and secondary data on the individual level offers great opportunities for using the strengths of different data sources synergistically and overcoming some of their limitations. Statutory health insurance data

‡ Geteilte Erstautorenschaft (Arnold/Druschke)

and medical data from kindergarten- and school-entry-examinations of Saxon health authorities are examples of already existing data sources that can complement cost-consuming primary data collections by valuable data sets and open up opportunities for longitudinal analysis.

ZUSAMMENFASSUNG

Ziel der Studie Das Datenlinkage von Primär- und Sekundärdaten erfreut sich in der Versorgungsforschung zunehmender Beliebtheit, birgt jedoch unter anderem in Bezug auf den Datenschutz einige Herausforderungen für die Beteiligten. Ziel der vorliegenden Arbeit ist es, das im Rahmen einer Kohortenstudie aus dem Bereich pädiatrischer Versorgungsforschung (EcoCare-PIn) angewandte methodische Vorgehen beim Linkage dreier Datenquellen darzulegen sowie praxisrelevante Erfahrungen zu berichten. Hierbei wird besonders auf notwendige regulatorische Maßnahmen bezüglich des Datenzuganges und -linkage sowie auf die Datenvalidierung zur Absicherung einer fehlerfreien Verlinkung eingegangen.

Methoden Unter Berücksichtigung aller datenschutzrelevanten Erfordernisse wurde auf individueller Ebene ein Linkage von a) pseudonymisierten Abrechnungsdaten einer gesetzlichen Krankenkasse zu in den Jahren 2007 bis 2013 geborenen Kindern aus Sachsen, b) Primärdaten einer postalischen Befragung von Eltern/Betreuern und c) medizinischen Daten der Kindergarten und Schuleingangsuntersuchungen sächsischer Gesundheitsämter durchgeführt. Das Grundprinzip des Datenlinkage-Konzeptes war die strikte Trennung der Stellen der Datenerhebung und Datenanalyse, was durch die Einrichtung einer Vertrauensstelle realisiert wurde.

Ergebnisse Herausforderungen betrafen insbesondere die umfangreichen regulatorischen Maßnahmen im Vorfeld des Datenzuganges sowie auch Datenschutzerfordernisse während der eigentlichen Studiendurchführung. Weiterhin erforderten technische Aspekte sowie die Datenvalidierung besondere Aufmerksamkeit und Ressourcen. Es wurden zahlreiche Validierungsschritte angewandt, um fehlerhaftes Datenlinkage zu vermeiden und die hohe Qualität des finalen Datensatzes zu sichern. Die Validierung beinhaltete sowohl Plausibilitätsprüfungen innerhalb der Primärdaten als auch Konsistenzprüfungen bezüglich Angaben, die sowohl in Primär- als auch Sekundärdaten vorhanden waren.

Schlussfolgerung Das individuelle Linkage von Primär- und Sekundärdaten eröffnet wertvolle Möglichkeiten, die Stärken verschiedener Datenquellen synergistisch zu nutzen und einige ihrer Schwächen zu kompensieren. Krankenkassendaten und Daten der Kindergarten- und Schuleingangsuntersuchungen sächsischer Gesundheitsämter stellen Beispiele für bereits vorhandene Datenkörper dar, die kostenintensive Primärdatenerhebungen um wertvolle Datenbestände ergänzen können und Möglichkeiten für längsschnittliche Analysen eröffnen.

Introduction

Scientific research strongly depends on the nature and the extent of the available data. By using questionnaires it is possible to collect information specially tailored to the research question, but on the other hand primary data collection is cost consuming and may be affected by different kinds of bias. Selection bias, recall bias or social desirability bias are examples [1–3]. Secondary data constitute an alternative data source. Being data which already exist, using them bears the potential to be very resource-efficient. Besides the potential cost efficiency due to the way of data collection, secondary data offer additional advantages, depending upon the respective provenience. Secondary data of German statutory health insurances (SHI data), for example, are available with a high degree of timeliness and completeness. Furthermore, longitudinal analysis is possible [4]. However, as secondary data have been collected for other purposes, they might not include all relevant information. In German SHI data, for example, sociodemographic and psychosocial information are limited. In addition, as the purpose of SHI data is billing of services, data may be subject to bias due to upcoding [5]. All in all, there are strengths and limitations with all different kinds of data sources. Thus, linking different data sources on the individual level offers great opportunities for using the strengths of different data sources synergistically and overcoming some of their limitations (e. g. [6]). The number of research projects that use linkage of different secondary and primary data to answer research questions increases (examples lidA-Kohortenstudie [7], "Gesundes Kinzigtal" [8]). The enormous challenge this contains is to realize linkage whilst safeguarding data privacy.

This paper will provide insight into the methods used in the linkage study EcoCare-PIn (Early comprehensive Care of Preterm Infants). This cohort study investigated the short- and long-term consequences of preterm birth with regard to parental stress, parent-child relationship, family and child quality of life, child development, and healthcare utilization including costs. For the investigation of this broad spectrum of clinical, psychosocial and socioeconomic outcomes, data linkage on individual level is a promising approach. While complying with all relevant data protection requirements, the study realized an individual-level linkage of

(1) pseudonymized **administrative SHI data** from a Saxon statutory health insurance on children born between 2007 and 2013 in Saxony (inpatient and outpatient data),
(2) **primary data** collected from the parents/caregivers of all eligible very low birth weight (< 1500 g) and low birth weight infants (1500–2500 g) and a matched sample of infants above 2500 g birth weight,
(3) **medical data from kindergarten- and school-entry-examinations** of Saxon health authorities (in extension of the EcoCare-PIn study).

SHI data and medical data from kindergarten- and school-entry-examinations of Saxon health authorities are two examples of already existing data sources that can complement cost-consuming

primary data collections with valuable data sets and open up opportunities for longitudinal analysis.

Using the study EcoCare-PIn as an example, the aim of this article is to systematically outline the methods used and the experiences made during the conduct of a linkage study in the field of pediatric health care research that involved access to and linkage of three different data sources, among them medical data from health authorities, which have previously hardly been used for research [9].

The article addresses necessary regulatory measures, data access, data validation and other preconditions for the successful and data protection-compliant data linkage and discusses challenges and possible solutions. From the experiences made during EcoCare-PIn, recommendations will be derived that may help researchers in planning and executing similar studies in (pediatric) healthcare research.

Methods

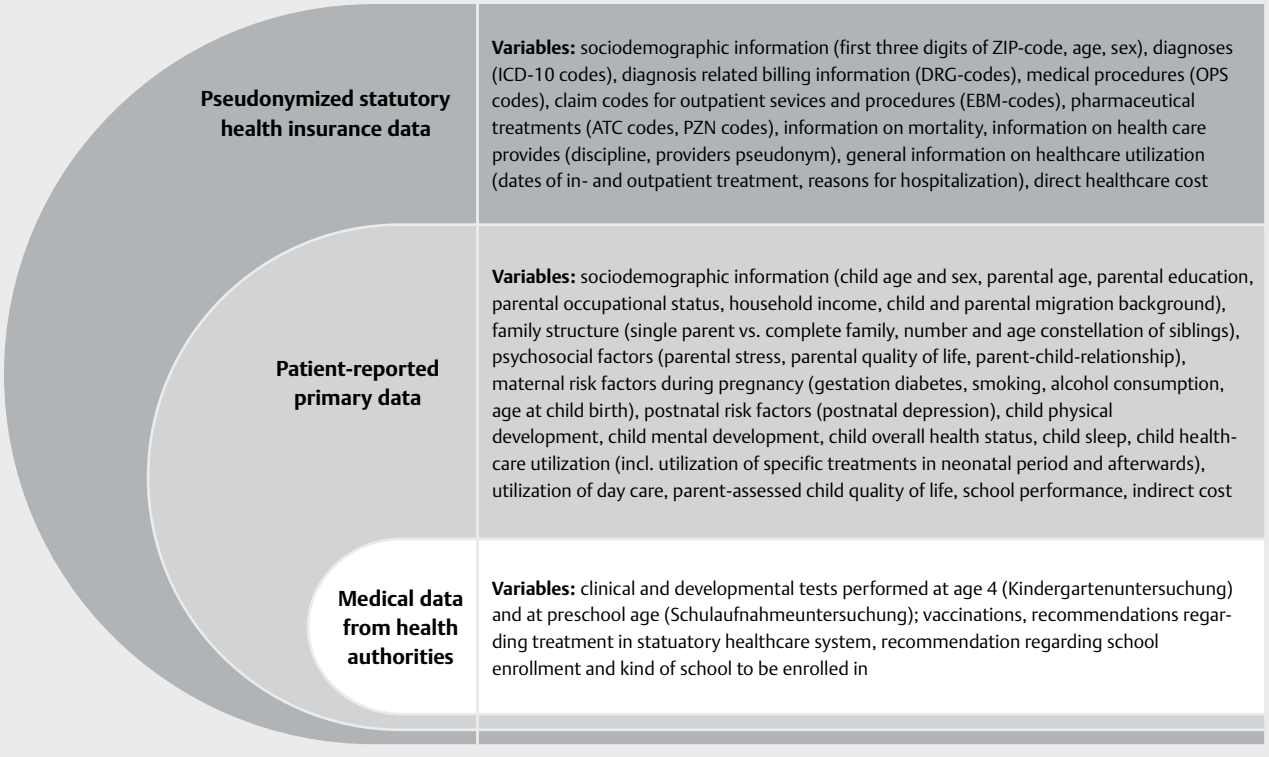
Study design, data sources and cohorts

The publicly funded cohort study EcoCare-PIn has been registered (Deutsches Netzwerk Versorgungsforschung: VfD_EcoCare-PIN_13_003463) and described elsewhere [10]. The methods and procedures of the study were developed in compliance with the ethical principles of the Declaration of Helsinki [11] and the guidance provided in Good Epidemiologic Practice [12] in connection with Good Practice of Secondary Data Analysis [13]. Standard Operating Procedures for handling of secondary data guaranteed the fulfilment of best practice requirements within the Center for Evidence-Based Healthcare (ZEGV) as data analysis site. The study protocol has been approved by the responsible ethics committee (EK 67022014) and the data protection officer of the Technische Universität Dresden as well as the Saxon data protection officer (2-7410-74/1) and the data protection officer of the SHI.

Data from three different sources were used and linked (see ▶ **Fig. 1**):

(1) First, the study was based on pseudonymized **administrative SHI data** that were provided by the German statutory health insurance AOK PLUS for all insured children within the Federal State of Saxony who were born between January 1st, 2007 and December 31st, 2013 as shown in ▶ **Fig. 1**. First results of the SHI data analysis have been published recently [14].

(2) To enable further analyses, additional **primary data** were collected on a subgroup of the described administrative insurance data cohort of children. This subgroup of children was selected as follows: Primarily, all not deceased children continuously insured from their birth onwards with a Saxon ZIP-code were defined as being eligible for primary data collection. Out of these, 17,500 children were to be selected for primary data collection (determination from sample size estimation). Since the project focused on the long term consequences of low birth weight, all children with very low and low birthweight were chosen. Control children for the primary data collection were



▶ **Fig. 1** Architecture and variables provided in EcoCare-PIn.

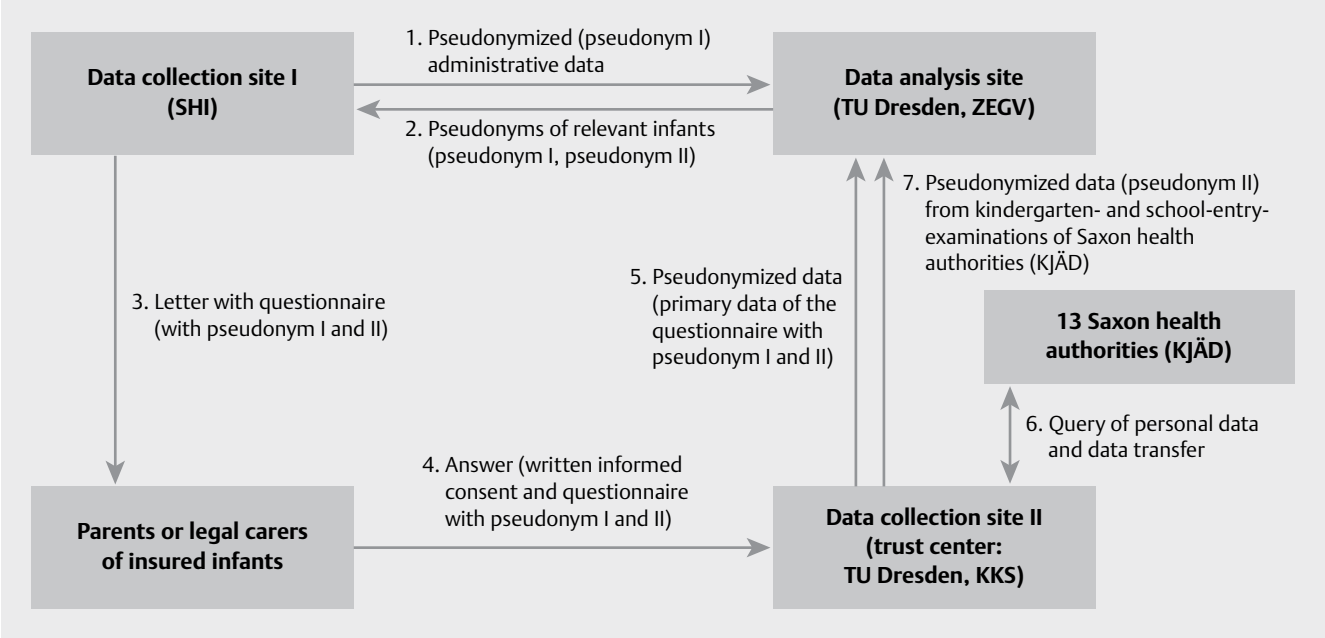
selected via frequency matching according to birth year, sex and administrative district. The contents of the primary data collection (postal questionnaire) are shown in ▶ **Fig. 1**.

(3) On a subgroup of children selected for primary data collection (ages 3 and up), there is availability of physician-collected medical data from kindergarten- and school-entry-examinations of Saxon health authorities. As an extension of the EcoCare-PIn study, these **standardized medical data from Saxon health authorities** were used. All 13 Saxon public health authorities conduct standardized clinical and developmental tests (fields of vision, hearing, language, fine and gross motor skills) of 4-and 6-year old children and document the results digitally via using an unitary software. While the school-entry-examination (at age 6) is compulsory for all children of the appropriate age, the kindergarten examination (at age 4) introduced in 2003 is voluntary, based on parental consent [15]. Therefore, the data source of Saxon health authorities contains files on all 6-year-old and on a share of 4-year-old children. Due to the update of records, longitudinal data are available for some children.

Concept of data linkage

A concept of data linkage (▶ **Fig. 2**) was developed to realize the individual-level linkage of the three mentioned data sources in accordance with data protection requirements. The fundamental principle of the concept of data linkage, which will be outlined below, was to strictly separate the sites of data collection (I and II) from the data analysis site (ZEGV).

1) Starting point was the site of data collection I (SHI), which submitted pseudonymized (pseudonym I) administrative outpatient and inpatient data of all children born in Saxony between January 2007 and December 2013 with health insurance at the AOK PLUS to the data analysis site (ZEGV).
2) Data analysis site (ZEGV) then identified individuals eligible for the primary data collection and transferred pseudonyms (pseudonym I) of all children with very low birth weight and low birth weight as well as of matched control children (children with normal birthweight and without recorded birthweight) to the site of data collection I (SHI). For reasons of better practicability (see below section C technical aspects), ZEGV additionally transferred a new short pseudonym per individual (pseudonym II).
3) Data collection site I (SHI) de-pseudonymized these information and sent postal questionnaires with pseudonym I and II to the caregivers of the selected children insured at the AOK PLUS. The questionnaires were supplemented by in-depth information about the study concept and data protection issues (study information) and an informed consent form for study participation and one (primary data and SHI data) or both (primary data and SHI data and health authorities data) linkage actions. Additionally, the carer(s) of all children aged 3 years or older were asked to give informed consent regarding the transfer and linkage of data from kindergarten- and school-entry-examinations of Saxon health authorities.
4) Children whose caregivers agreed with study participation sent the completed questionnaire and the signed consent form to the data collection site II (trust center Koordinierungszentrum für Klinische Studien Dresden (KKS)).
5) Data collection site II (trust center) received the completed questionnaires, deleted any personal identifiers and sent the pseudonymized datasets to the data analysis site (ZEGV) for the linkage with health insurance data and statistical analysis.
6) Furthermore, data collection site II (trust center) realized the personal query of the medical data of health authorities using the pseudonym for those children with written informed consent. Initially, the health authorities had been provided with a list of contemplable children and copies of the respective writ-



▶ **Fig. 2** Concept of data linkage.

ten informed consent forms by the trust center. Then the responsible person of the data collection site II (trust center) asked for the respective datasets in the premises of the health authorities where they were filtered out and pseudonymized. Only pseudonymized datasets left the health authorities.

7) Data collection site II (trust center) electronically provided the pseudonymized datasets to the data analysis site (ZEGV) for the linkage with health insurance data and questionnaire data.

Linkage was possible by continuously using pseudonym II, without personal identifiers, keeping all data protection requirements in mind. As a result of the concept of data linkage, throughout the study only pseudonymized primary data and/or pseudonymized secondary data were transmitted to the data analysis site. Solely the trust center received personal data of the persons involved. The trust center, in turn, had no access to the SHI data. Once the parties had given their consent and the completed questionnaires were sent to the trust center, this information was forwarded to the ZEGV for data analysis only with a pseudonym. Secondary data and personal identifiers were not in one hand outside the SHI at any time.

Results

Response rates

EcoCare-PIn drew upon pseudonymized SHI data on 139,383 Saxon children born between 2007 and 2013 (▶ **Fig. 3**). From this data pool, 17,498 datasets were selected by the data analysis size (ZEGV) including all very low and low birthweight children as well as a matched sample of control children. The pseudonyms of these children were transferred to the health insurance. Due to missing addresses, death or the lack of a contact permission, the health insurance did not contact the carer(s) of 985 children. The health insurance approached the carer(s) of the remaining 16,513 children with a postal questionnaire carrying the pseudonyms, asked them to fill it in and to give informed consent regarding a linkage of questionnaire and health insurance data. The responses of the primary data collection (adjusted response rate: n = 4,512 (27.3 %)) were received by the trust center. Additionally, the carer(s) of 2,627 children aged 3 years or older (equals carers of almost 90 % of children at this age) gave informed consent regarding the transfer and linkage of data from kindergarten- and school-entry-examinations of Saxon health authorities. Finally, for 1,677 of these children data were transferred from the 13 Saxon health authorities. The difference between the number of children with informed consent and the number of transferred datasets arised from two causes: either the children were not yet four years old and therefore had not yet been due for the kindergarten examination or they were already four years old but no kindergarten examination had taken place (voluntary examination depending on parental consent; sometimes also suspended by health authorities in favour of conducting the obligatory school-entry-examinations). The trust center sent the pseudonymized questionnaire data and the health authorities' data to the data analysis site.

Challenges and approaches

A) Regulatory measures in advance to get data access

Comprehensive preliminary considerations and regulatory measures in advance were necessary to get data access whilst safeguarding data privacy. Several responsible data protection officers had to be engaged to prove the overall concept of the study including all study documents. The approval of the study was based on the concept of data linkage (▶ **Fig. 2**), data protection- and data security concepts of the institutes involved as well as on study documents for participants of the primary data collection.

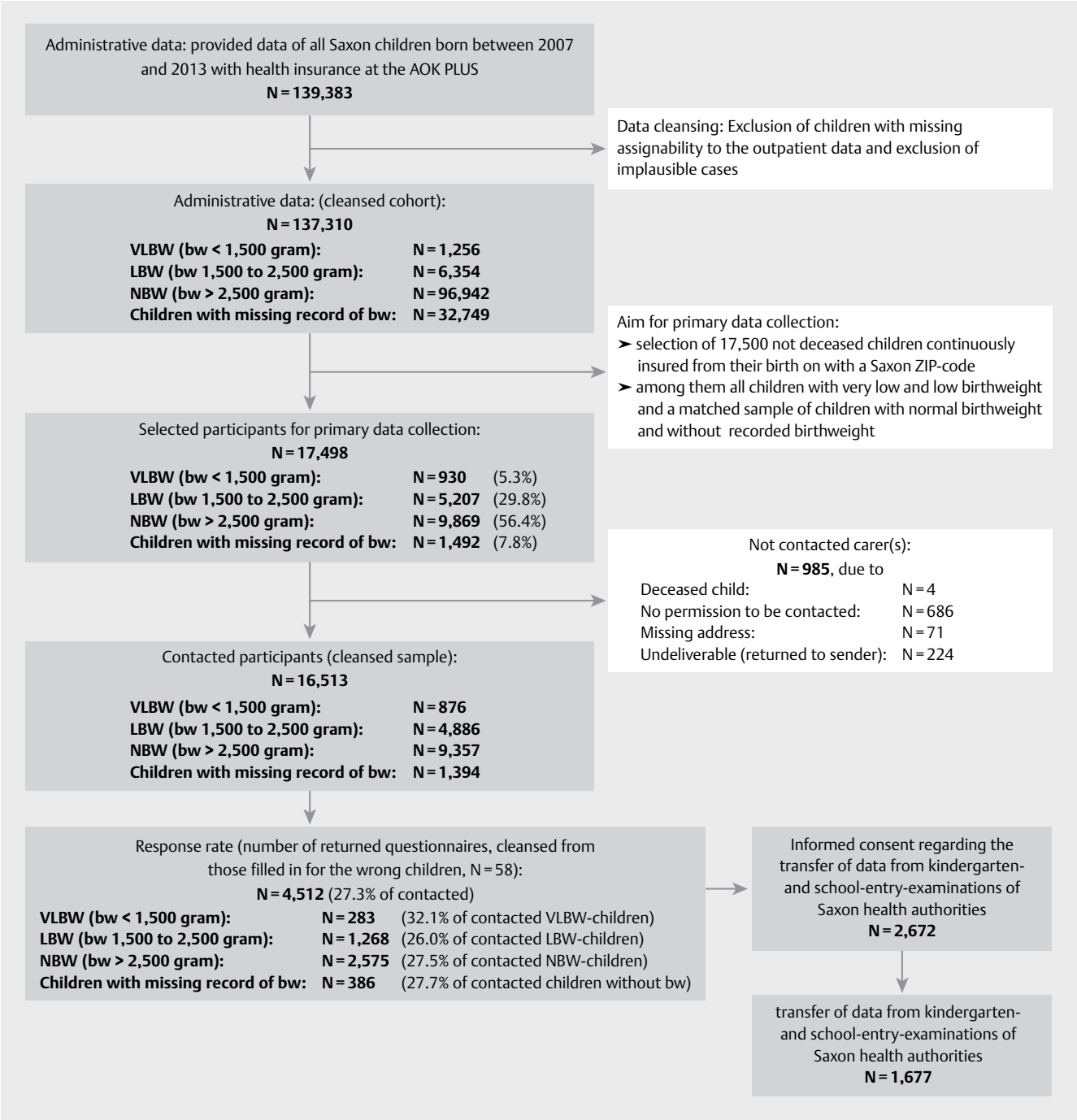
(1) In EcoCare-PIn, **access to SHI data** of the health insurance was contract-based and took place on the basis of a detailed data record description. The health insurance participated as project partner, since there was strong interest in the assessment of the burden of preterm birth. A positive internal evaluation of the data protection officer of the health insurance (data owner) was performed regarding the study on the basis of the legislation valid at the time of data transfer. The SHI comprehensively examined the necessity of an application according to §75 of the German Social Code, book X (SGB X). At the time of preparing data access, this application was deemed not necessary by the SHI. The de-pseudonymization of the pseudonymized SHI data was regarded impossible outside the SHI and data therefore considered "de facto anonymous". According to §3 VI of the Federal Data Protection Act old version (German: BDSG a.F.), anonymization meant that data can either not be attributed to an individual or only with a disproportionate amount of time, expense and labour. However, with the commencement of the General Data Protection Regulation (GDPR, German: Datenschutzgrundverordnung DSGVO) on 25 May 2018, the latter mitigation is not provided any more [16]. Methodological guidance for the realization of data linkage with regard to the current data protection regulations is published in Good Practice Data Linkage [17].

(2) The regulatory measures to get **access to primary data** concern a) access to the individuals for primary data collection and b) the collection and linkage of the data themselves.

Access to **individuals for primary data** collection was realized via the SHI who postally approached the caregivers of eligible children with the following documents:

- in-depth information about the study concept and data protection issues (study information)
- questionnaire for primary data collection
- consent form (2x) regarding (1) participation in primary data collection and data linkage with SHI data and (2) regarding the transfer of data from kindergarten- and school-entry-examinations of Saxon health authorities and their linkage with SHI and primary data
- nonresponse-form
- small incentive (colouring picture for children)

For reasons of data protection, the transfer of information about participation or non-participation of insured persons to the SHI had to be ruled out. Therefore, the contribution of the SHI ended with sending the study documents to the caregivers of eligible children. To collect the returning questionnaires, a trust center was delegated (▶ **Fig. 2**).



▶ **Fig. 3** Data flow.

Primary data collection and linkage was legitimized through obtaining written informed consent from the children's caregivers for study participation and one (primary data and SHI data) or both (primary data and SHI data and health authorities data) linkage actions. In case of participation in the primary data collection, the initially "de facto anonymized" SHI data had to be regarded as personal data, whose scientific use was to be covered by informed consent.

(3) Written informed consent was also the legal basis for the **use of the health authorities' data.** The head officials of the Saxon health authorities were postally asked to support the study EcoCare-PIn by providing medical data from kindergarten- and school-entry-examinations on children whose caregiver(s) had documented their consent. Considerable effort was made to motivate health authorities (individual invitation, covering letter of the Saxon State Ministry of Social Affairs, project presentation). Together with IT personnel, resource-saving procedures for data extraction and transfer were planned and the health authorities were offered various options of receiving assistance from the trust center (KKS). Nonetheless, the support of the

project meant an additional burden for the health authorities.

Still, all 13 health authorities could be motivated to participate. Overall, the **involvement of a trust center** was an absolutely necessary regulatory precondition for the conduct of data collection and linkage. As an organizationally, spatially and personally independent unit [18], it enabled the collection of primary data and health authorities data and their pseudonymized allocation to the data analysis site.

B) Data protection requirements during the study

To increase the response rate of the primary data collection, a reminder letter was sent after four weeks. Due to data protection requirements, it was not allowed to transfer the information to the SHI which children's caregivers had already answered and returned the questionnaire. Therefore, the SHI had to repeatedly send the questionnaires (again supplemented by in-depth information about the study concept and data protection issues and a consent form) to the whole subcohort of individuals selected for the primary data collection.

Despite a note in the reminder letter, that the added questionnaire please be ignored if the initial one had been returned, some participants filled in and sent the questionnaire a second time. In this case, the more complete one was used for the analysis.

C) Technical aspects

Transfer of health insurance data to the data analysis site (ZEGV)

SHI data were provided via a portal for secure data transfer.

Data extraction in health authorities

Considerable effort was made to clarify with IT personnel of health authorities how data extraction is technically possible. As most health authorities had sparsely or not been involved in research projects, there were no routine procedures implemented yet. However, as all authorities used the same data software and were managed by the same IT company, communication could initially be centralized and a unified technical solution for data extraction could be established. This solution was then communicated to and applied by IT managers of each single health authority. After data extraction in the health authorities, the data had to be transferred to the data collection site II (trust center) in line with data protection requirements. Pseudonymized data were personally collected by the responsible person of the trust center in the form of password-protected CD's.

Electronical reading of questionnaire data

Due to the sample size of EcoCare-PIn, electronic reading of questionnaire data was opted for. From previous research projects, the reading software was known to have some difficulty with distinguishing between some letters and numbers, i. e. between O and 0 or between 1 and 7. To minimize sources for reading mistakes, a second pseudonym (pseudonym II), comprising of only five numbers, was assigned to every questionnaire by the data analysis site in addition to pseudonym I, which had been assigned by the SHI and was a combination of 24 numbers and letters. Only the short pseudonym II was used for the electronic reading of questionnaires.

Transfer of primary and health authorities data to the data analysis site (ZEGV)

Data collected by the trust center were provided via internal network.

D) Validation of data and data linkage

Considerable effort was spent on data validation in order ensure the high quality of the final dataset and avoid incorrect data linkage. A number of validation routines were applied. The validation focused both on plausibility checks within the questionnaire data (e. g. check whether answers are within valid range of values, comparison of different variables related to the age of the child) and on consistency checks of information given in primary and secondary data (i. e. birthweight and age-related information). In advance, tolerance ranges had been determined together with pediatric clinicians. For example, differences in birthweight between primary and secondary data were determined tolerable if ≤ 100 g. Larger differences meant that primary and SHI data were not linked.

Regarding the comparison of birthweights in primary and secondary data, many initial "discrepancies" could be eliminated through simple correction, as their cause was clearly traceable (see ▶ **Table 1**): the parental confusion of questionnaires between siblings within one family, transposed digits in hand written birthweight in questionnaires or reading errors of the program regarding hand written birthweight. Furthermore, some "discrepancies" in birthweight were caused through the fact that – in the case of healthy newborn multiples – the SHI assigned the birthweight of one sibling identically to the other sibling(s).

After having identified and corrected all of these "false alarms", there remained 58 questionnaires which were obviously not filled in for the correct child and therefore were not used for data linkage.

To validate the linkage of the different data sources several variables (e.g. age) were used that are concordantly recorded in all data bodies to be linked.

Discussion

The present article outlined the methodological approach used in the linkage study EcoCare-PIn. The experiences of linkage on an individual level within the study shows the feasibility of integrating data from the health authorities and may serve as a blueprint for subsequent studies which want to link data from various data sources individually. Linkage is possible and worthwhile to create a solid database for the further development of perinatal healthcare in Germany. Derived from the experiences of EcoCare-PIn, the following points summarize the main insights that may be of interest for researchers who intend to conduct similar studies:

The complexity of the data-flow concept and of the validation process increases with increasing number of data sources that are aimed to be linked. The providers of secondary data and their data protection officers therefore need to be involved from an early stage.

There was considerable effort to motivate all 13 health authorities to participate, since the project meant an additional workload. Therefore, supporting them with the implementation of a data extraction routine was highly appreciated. Special structural institutions (trust center) are indispensable to link the data sources on individual level in accordance with data protection requirements.

The complexity of the study also led to high requirements in the preparation of study documents to describe the procedure in a comprehensible manner for the participants. A written informed

▶ **Table 1** Patterns of discrepancies. The table summarizes the different patterns of discrepancies found while checking the consistency of information given in primary and SHI data. There might be an overlap within the frequencies.

Types of discrepancies	Examples	Approach	Causes	Prevention	Frequency
1. Birthweight information of two questionnaires with adjacent serial numbers do not fit to birthweight information in SHI data, the questionnaires seemed to be interchanged	Parents wrote on questionnaires that they confused the questionnaires	Data (pseudonym 2) was assigned correctly and the data were linked (no further differences)	Parents confused the questionnaires for siblings	Name target child in cover letter and study information	n = 36
2. Reading error of the program regarding the handwritten birthweight of the primary data collection	Birthweight in SHI data: 3,430 g Birthweight in the completed questionnaire: 8,430 g	Primary data have been corrected and the data were linked (no further differences)	Reading error with critical numbers (1 and 7, 5 and 8, 3 and 8) or indistinct spelling	Visual inspection during/after the reading process necessary	n = 8
3.Transpositions in the hand written birth weight information given in the primary data collection	Birthweight in SHI data: 2,240 g Birthweight in the completed questionnaire: 2,420 g	Primary data have been corrected and the data were linked (no further differences)			n = 20
4. Accumulation of discrepancies of birthweight-information in questionnaire vs. SHI data for twins/muliples	Birthweight of both children in SHI data: 2,200 g Birthweight in the completed questionnaire: 2,560 g and 2,200 g	Primary data were used and the data were linked	In SHI data in case of healthy twins/multiples the same birthweight (of one of them) was allocated to all multiples since healthy newborns (and their birthweight) are coded together with the mother's delivery.		n = 29
5. Discrepancies of birthweight/age-information in questionnaire vs. SHI data, that cannot be clarified	Age of 4 years in SHI data vs. 7 years in the completed questionnaire Birthweight of 2,450 g in SHI data and 2,600 g in the completed questionnaire	Age: Evaluation was implausible, data were not linked Birthweight: consultation of clinical experts → differences of maximal 100 g were defined to be acceptable, children with larger differences were excluded Of 4,074 children both birthweight-information exist (SHI and and primary data) no difference in both data sources: 82.8 % differences of maximal 100 g: 11.5 % differences of more than 100 g: 5.7 %	Birthweight: DRG-Upcoding [15] Wrong child, parents confused the questionnaires for siblings Wrong child, i. e. care givers of children selected as control child from the NBW group coincidentally also had a preterm child within the family and therefore filled in the questionnaire for the preterm child	Name target child in cover letter and study information Slightly modified study information for the control group	n = 58

consent was the legal basis for usage and linkage of the secondary and primary data. The study highlights the value of data from health authorities, which have so far received little attention from research [8].

When selecting participants for the survey from SHI data it should be taken into account that - due to missing addresses, death or the lack of a contact permission - the health insurance may not be able to contact the whole target population. In the case of Eco-Care-PIn, the SHI did not contact the carer(s) of 985 children. This is a considerable loss of a substancial group of children chosen for primary data collection that is necessary to be taken into account in sample size estimation and also in considerations regarding the generalizability.

The resending of all study documents at the reminder, which was important for data protection reasons, caused confusion and displeasure among some participants and moved some participants to repeatedly fill in the questionnaire. This resulted in an increased effort in the documentation of the return and in the validation process.

The existence of several variables that are concordantly recorded in all data bodies to be linked proved useful for validation purposes. However, in case of discrepancies, decisions have to be made regarding the degree of data non-conformance that is still acceptable or is considered to indicate incorrect data linkage. Thus, data validation is time consuming and has to be scheduled in advance. For the definition of tolerance ranges, interdisciplinary exchange with clinicians and data providers is indispensable.

Recommendations for practice

- schedule enough time in advance to clarify data access and establish/involve trust center
- select a sufficiently high number of SHI patients based on sample size estimations, taking into account that probably not all insurants can be contacted by the SHI
- if possible: collect several variables via primary data collection although they are also available in secondary data (for validation purposes)
- schedule enough time for data and data linkage validation
- include clinicians to determine reasonable tolerance ranges for data validation

Acknowledgements

This study was funded by the Federal Ministry of Education and Research (BMBF, 01GY1323). With regard to the data provision for EcoCare-Pln we thank the AOK PLUS, the Saxon health authorities including their IT support company easysoft GmbH Dresden and the participants of the primary data collection. We also thank the KKS (trust center) and Dr. Enno Swart (external quality assurance) for their indispensable support of the study.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Hunger M, Schwarzkopf L, Heier M et al. Official statistics and claims data records indicate non-response and recall bias within survey-based estimates of health care utilization in the older population. BMC Health Services Research 2013; 13: 1. doi:10.1186/1472-6963-13-1

[2] Janssen C, Swart E, von Lengerke T., Eds. Health Care Utilization in Germany. Theory, Methodology and Results. New York: Springer; 2014

[3] Peersman W, Pasteels I, Cambier D et al. Validity of self-reported utilization of physician services: A population study. European Journal of Public Health 2013; 24: 91–97. doi:10.1093/eurpub/ckt079

[4] Swart E, Stallmann C, Schimmelpfennig M et al. Gutachten zum Einsatz von Sekundärdaten für die Forschung zu Arbeit und Gesundheit. 1. AuflageDortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedi-zin; 2018. doi:10.21934/baua:bericht20180112

[5] Jürges H, Köberlein J. What explains DRG upcoding in neonatology? The roles of financial incentives and infant health. Journal of Health Economics 2015; 43: 13–26. doi:10.1016/j.jhealeco.2015.06.001

[6] Swart E, Thomas D, March S et al. Erfahrungen mit der Datenverknüpfung von Primär- und Sekundärdaten in einer Interventionsstudie. [Experience with the linkage of primary and secondary claims data in an intervention trial]. Gesundheitswesen. 2011; 73: e126–e132. doi:10.1055/s-0031-1280754

[7] March S, Rauch A, Thomas D et al. Datenschutzrechtliche Vorgehensweise bei der Verknüpfung von Primär- und Sekundärdaten in einer Kohortenstudie: Die lidA-Studie. Das. Gesundheitswesen 2012; 74: 834–835. doi:/10.1055/s-0031-1301276

[8] Schulte T, Pimperl A, Dittmann B et al. Drei Dimensionen im internen Vergleich: Akzeptanz, Ergebnisqualität und Wirtschaftlichkeit der Integrierten Versorgung Gesundes Kinzigtal. 2012; Im Internet: http://deutsche-aerztenetze.de/uploads/live/aktuelles/dokumente/24/studie_kin-zigtal.pdf Stand: 22.11.2018

[9] Weyers S, Wahl S, Dragano N et al. Ist der Datenschatz schon gehoben? Eine Übersichtsarbeit zur Nutzung der Schuleingangsuntersuchung für die Gesundheitswissenschaften. Prävention und Gesundheitsförderung 2018, doi:10.1007/s11553-018-0641-6

[10] Schmitt J, Arnold K, Druschke D et al. Early comprehensive care of preterm infants - effects on quality of life, childhood development, and healthcare utilization: study protocol for a cohort study linking administrative healthcare data with patient reported primary data. BMC Pediatrics. 2016; 16: 104. doi:10.1186/s12887-016-0640-8

[11] World Medical Association World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. JAMA 2013; 310: 2191–2194. doi:10.1001/jama.2013.281053

[12] Deutsche Gesellschaft für Epidemiologie (DGEpi). Leitlinien und Empfehlungen zur Sicherung von Guter Epidemiologischer Praxis (GEP). 2008. Im Internet: https://dgepi.de/assets/Leitlinien-und-Empfehlungen/6074a4e7b8/Leitlinien-fuer-Gute-Empidemiologische-Praxis.pdf Stand: 13.11. 2018

[13] Swart E, Gothe H, Geyer S et al. Good Practice of Secondary Data Analysis (GPS): Guidelines and recommendations. Gesundheitswesen 2015; 77: 120–126. doi:10.1055/s-0034-1396815

[14] Rüdiger M, Heinrich L, Arnold K et al. Impact of birthweight on health-care utilization during early childhood – A birth cohort study. BMC Pediatrics 2019; 19: 69. doi:10.1186/s12887-019-1424-8

[15] Freistaat Sachsen. Gesetz über Kindertageseinrichtungen in der Fassung der Bekanntmachung vom 15 Mai 2009 (SächsGVBl. S. 225), das zuletzt durch Artikel 7 des Gesetzes vom 29. April 2015 (SächsGVBl. S. 349) geändert worden ist. Im Internet: www.revosax.sachsen.de Stand: 17.05.2018

[16] Arbeitsgruppe „Datenschutz und IT-Sicherheit im Gesundheitswesen“ der deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS). Arbeitshilfe zur Pseudonymisierung/Anonymisierung. 2018; http://ds-gvo.gesundheitsdatenschutz.org/download/Pseudonymisierung-Anonymisierung.pdf Stand: 13.11.2018

[17] March S, Andrich S, Drepper J et al. Gute Praxis Datenlinkage (GPD). Gesundheitswesen. 2019, doi:10.1055/a-0962-9933

[18] Pommerening K, Drepper J, Helbing K et al. Guideline for Data Protection in Medical Research Projects – TMF's generic solutions 2.0. TMF-Book Series. Vol.11. Berlin: MWV; 2014