# BMJ Open

# Data linkage of German statutory health insurance claims data and care needs assessments preceding a population-based cohort study on nursing home admission

Dominik Domhoff [1,2] Kathrin Seibert [1,2] Susanne Stiefler [1,2] Karin Wolf-Ostermann [1,2] Dirk Peschke[3]

¹Institute of Public Health and Nursing Research, University of Bremen, Bremen, Germany
²High Profile Area Health Sciences, University of Bremen, Bremen, Germany
³Department of Applied Health Sciences, Hochschule für Gesundheit Bochum, Bochum, Germany

**Correspondence to**
Dominik Domhoff;
ddomhoff@uni-bremen.de

## ABSTRACT

**Objectives** We perform and evaluate record linkage of German Care Needs Assessment (CNA) data to Statutory Health Insurance (SHI) claims data. The resulting dataset should enable the identification of factors in healthcare predicting the time between the onset of long-term care dependency and the admission to a nursing home in Germany in subsequent analyses.

**Design** A deterministic record linkage was conducted using the key variables region, sex, date of birth and care level. In further steps, the underlying cause of care dependency (International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10)) was added for a higher level of distinction. Before linkage, the suitability of the two datasets for these procedures was assessed. After linkage, the results of each stage were analysed and the resulting dataset was evaluated cross-sectionally with respect to bias generated through this process.

**Setting** The study comprises data from the German SHI and Statutory Long-Term Care Insurance.

**Participants** The study cohort comprised 158 069 individuals who became care dependent in 2006. We obtained CNA data for the year 2006 including 188 935 individuals.

**Results** We could link CNAs to 66 310 individuals of the original study cohort, corresponding to 42.0%. Records from two federal states could not be matched due to missing data. Linkage rates were lower where more people shared the same attributes. The resulting dataset showed minor differences regarding age, sex and care level compared to the original cohort.

**Conclusions** Data linkage between German SHI claims data and CNA data is feasible. Failure to link was mostly attributable to a lack of distinction between individuals using available identifiers. The resulting dataset contains relevant information from both health services provision and functional status of care dependent people and is suitable for further analyses with critical reflection of representativity.

## BACKGROUND

The admittance to a nursing home is deemed undesirable by the affected: care dependent people prefer to live in a setting with a high degree of personal autonomy, which is preferably their traditional home.[1–4] Considering the individuals' preferences is again expected to be beneficial in terms of healthcare outcomes.[1] While knowledge on preferences is particularly important for the planning of long-term care and housing arrangements themselves, efforts towards the prevention of institutionalisation of care dependent people need to be advanced.

Besides the impact of the home care situation, the health status is found to be an important predictor of institutionalisation of a care dependent person. Cognitive impairments and functional impairments showed stronger evidence for predicting nursing home admission while the presence of single diagnoses like stroke, hypertension, respiratory diseases or arthritis was not conclusive regarding their impact on institutionalisation.[5]

In Germany, insurances for healthcare and long-term care are compulsory. Around 90% of the population is insured through Statutory

---

**STRENGTHS AND LIMITATIONS OF THIS STUDY**

⇒ This study is one of the first linking German Statutory Health Insurance claims data to Care Needs Assessment data albeit the lack of a unique identifier.

⇒ A large and representative sample of insurees was included to perform and evaluate the linkage process.

⇒ Three-phase data linkage increased the quantity of linked records and the number of included and corresponding attributes.

⇒ Available routine data lacks information to reliably validate the accuracy of the performed data linkage.

Long-Term Care Insurance (LTCI) and Statutory Health Insurance (SHI) provided by insurance funds. Substitutive choice of private health and LTCI is limited, primarily based on employment status and income.[6] SHI and LTCI are separate, but largely provided by the same insurance funds and companies.

SHI provides in-kind services, including inpatient and outpatient medical care, prescription drugs, medical rehabilitation, physical and occupational therapy, based on medical needs, without capping, but following the principles of being 'adequate, appropriate and efficient'.[6] LTCI provides in-kind services and cash benefits to people in need of nursing care. A Care Needs Assessment (CNA) is performed on request and suspicion of long-term care dependency and when related changes in a person's condition occur. It can be initiated by anyone, requiring the consent of the affected person or their legal guardian. LTCI companies have delegated the conduction of CNA to the regional Medical Services of the German SHI providers (MDK). There are 15 regional MDK, which are associated in the Medical Advisory Service of the German Association of SHI Funds (MDS) that is coordinating their work on the national level. The assessment results have been used by the SHI funds to decide on the allowance of a care level between 1 and 3 until 2016, where classification was performed on the average time of care a person was expected to need per day. In 2017, CNA was reformed and eligibility of LTCI benefits was based on five care levels,[7] with individual abilities and resources playing a key role. The care level is granted retroactively to the date of application. The amount of benefits is capped per care level and the selected type of services (in-kind, cash or a combination of both).[7] Therefore, LTCI is not designed to necessarily cover all costs of care dependency.

Provision of long-term care services and its funding through LTCI is separate from the provision of health services and the respective funding through the SHI.[8] With the individuals SHI company regularly providing the LTCI, claims data for both sectors are available in one place. In-kind services in the SHI are billed with a high granularity based on diagnoses from the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10), procedure codes (Operation and Procedure Classification System for inpatient, Uniform Value Scale for outpatient care) and provided medication (Anatomical Therapeutic Chemical Classification). However, claims data from LTCI only provide very limited insights into care arrangements, as they only comprise care level, residency in a nursing home or community-dwelling and the amounts of in-kind services and cash benefits.

In contrast, CNA are usually performed in presence of the person in need of care and in their home. For the assessment, functional and cognitive status of the applicant is taken into account as well as housing, living and care arrangements.[9] Information gathered during CNA therefore provides an indispensable insight into conditions affecting their choice of living. Hence, these assessments combined with SHI claims data markedly increase the possibility of understanding the factors related to the time until institutionalisation.

The project 'Nursing Home Admission and its Predictors in Healthcare Quality, Living and Assistive Arrangements—a Population-based Cohort Study' aims to explore factors predicting the time between the onset of long-term care dependency according to the 11th book of the German Social Code (SGB XI) and the admission to a nursing home. These predictors include information regarding living and housing conditions and the functional status of the care dependent people as well as their inpatient and outpatient healthcare and rehabilitation history. As there is no comprehensive dataset comprising these information, we had to combine nationwide claims data from a major German SHI fund, the AOK, with data from the CNA, held by the MDS. As both datasets originate from different data holders and do not comprise a common unique identifier, linking the individuals' records for further analyses has to be conducted by combining several individual attributes.

This article describes the process of linking the two datasets without the presence of a single unique identifier. The aims are to (1) assess the comparability of both datasets, (2) perform and describe the data linkage, (3) assess validity and plausibility of the linkage results with special respect towards the introduction of selection bias into the linked dataset to (4) consequently obtain a dataset allowing to assess institutionalisation of care dependent people in regard to housing and living conditions as well as their received healthcare in subsequent analyses.[10–13]

## METHODS

This study presents cross-sectional results of a record linkage performed in preparation for a retrospective closed cohort study using secondary data, not primarily collected for scientific use. To achieve a maximum follow-up of 11 years until nursing home admission as the primary endpoint, we only included individuals in the study cohort living at home and becoming eligible for LTCI benefits in 2006. Individuals had to be at least 65 years old in this year. Two datasets were obtained from their respective data holders, which are also partners in the research consortium. The Scientific Institute of the AOK delivered claims data for all people in the study cohort for the years 2006–2016. The MDS provided the CNA data for the year 2006. The data protection officers of both institutions acknowledged the study including the linkage of the two datasets. We followed published recommendations for quality assurance[14] and reporting[15] of data linkage studies where applicable and aligned reporting with the Strengthening the Reporting of Observational Studies in Epidemiology statement for cross-sectional studies.[16]

## SHI claims data

SHI claims data of all individuals of at least 65 years of age, insured by one of the nationwide 17 German AOK insurance companies, who became care dependent and received benefits from the LTCI in 2006 amounted to 158 069 people and represent the study cohort. Only individuals insured at the AOK in every quarter of the year 2006, who survived the whole year and did not move into a nursing home in this year were included. Data contained the entire individual inpatient and outpatient health care history, including diagnoses and procedures, medication, rehabilitation services, physical therapy, occupational therapy, speech and language therapy, podology, type of long-term care benefit (home care or nursing home care) and personal data with federal state, postcode, date of birth and sex.

For linkage purposes, we used the personal data and care level in combination with ICD-10 diagnoses in 2006 and 2007. As there is evidence that especially coding quality of outpatient diagnosis in claims data may lack validity,[17–21] we only used those diagnoses considered of high validity: for inpatient diagnoses, we took into consideration ICD-10 codes diagnosed on discharge from hospital. Outpatient diagnoses were validated using the M2Q criterion.[20] This says that a diagnosis is only considered valid if it occurs at least once in the four quarters of the year following the initial diagnosis. Only diagnoses according to ICD-10 German Modification (ICD-10-GM) labelled as ascertained or status post were used for this procedure, diagnoses of exclusion and suspected diagnoses were omitted.

## CNA data

The provided data contained all results of CNAs of AOK insurants requested at one of the 15 regional MDK in 2006 and collected by the MDS according to the Statistical Guidelines of the German Association of LTCI Funds. For the states Lower Saxony and Bremen, AOK-insured could not be determined. Therefore, CNA data for all SHI-insured people, regardless of the insurance fund, were provided for these two states.

The CNA dataset contains information on needs for assistance in activities of daily living and the respective time required for assistance, available and recommended therapeutic appliances, cohabitation and persons aiding, recommendations for further services from LTCI and SHI and general information on the process of the CNA. Needs assessments are not only performed on first application for LTCI benefits but also occur as reassessments after the insured entered an objection or applied for a higher care level. Consequently, the dataset possibly contained multiple assessment records per person as well as records from people that received LTCI benefits prior to the year 2006. These records did therefore not match the inclusion criteria of the study and had to be excluded in a deduplication process.

In a first step, we identified reassessments of individuals with an initial assessment in 2006. As there was no

personal identifier in the CNA data, we had to identify records originating from the same person by using suitable variables in the available data. We sorted the data by region (14 attributes), regional SHI identifier (1531 attributes), sex (2 attributes), date of birth, occasion for assessment (3 attributes: initial assessment, reassessment after applying for higher care level, reassessment after objection) and date of assessment. Records were assumed to belong to a single person, when MDK region, SHI identifier, sex, and date of birth were equal, and a reassessment stated the same a priori care level as granted after a prior assessment.

Assessments from individuals for whom no initial assessment in 2006 could be determined were then excluded, as they did not match the study's inclusion criteria. Furthermore, we excluded all individuals less than 65 years old in 2006, individuals who were not granted a care level after assessment, individuals that had a care level before 2006, and individuals who applied for LTCI benefits for nursing home care in accordance with the project's inclusion criteria. We used information from only the last assessment in 2006 for consistency with the SHI claims data only including the latest information for that year. This resulted in 188 935 individuals CNA data records for further proceeding.

## Linkage

We identified four key variables common to both datasets for the purpose of a deterministic record linkage: (1) region, (2) sex, (3) date of birth and (4) care level. The needs assessment data additionally contain a single (5) underlying cause of care dependency coded as ICD-10. This attribute can be found in the inpatient and outpatient diagnoses of the SHI claims data, resulting in a fifth variable with a one-to-many relationship.

Attributes of these variables had to be transformed to a common denomination. This included condensing the regional attribute in the SHI claims data from 16 federal states to fit the 15 MDK regions. Hamburg and Schleswig-Holstein as well as Berlin and Brandenburg had to be merged, as each two federal states share one MDK. North Rhine-Westphalia has two separate MDK regions which were also merged, as information from SHI claims data did not provide this allocation. This resulted in 14 regions for the purpose of linkage. Furthermore, days and months of birth for observations from the states Bavaria and Baden-Wuerttemberg were missing in the CNA data and set to 1 January in both datasets. This should evade an otherwise necessary early exclusion of these observations and retain a possibility for a successful linkage.

As the underlying cause of care dependency was invalid or not filled in 12.6% of all assessments and the validity being uncertain as presenting medical diagnoses on assessment is not mandatory, we chose a stepwise approach. Since we expected the two datasets to contain identical individuals (with excess records from Bremen and Lower Saxony in the CNA data), which may not necessarily be distinguishable by the key variables, we added more
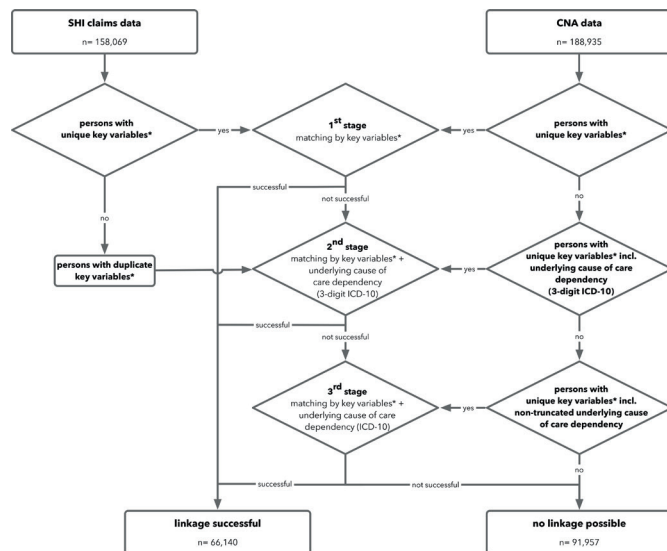
**Figure 1** Flowchart of the linkage process. *Key variables: region, sex, date of birth, care level. CNA, Care Needs Assessment; SHI, Statutory Health Insurance; ICD-10, International Statistical Classification of Diseases and Related Health Problems, 10th revision.

detailed variables in the following steps, when matches were not unique. The process is depicted in figure 1.

In a first stage, we extracted all observations from the respective datasets unique by the identifiers region, sex, date of birth and care level. Only if matches were not distinct, that is, one or more individuals in at least one of the datasets shared the same identifiers, we passed them on to a second stage.

In the second stage, we linked individuals not uniquely linked in the first stage or not distinct by region, sex, date of birth and care level by adding the underlying cause of care dependency truncated to a 3-digit ICD-10 as a linkage variable. Truncation was performed to account for our left-censored claims data with the possibility of minor variations in the available diagnoses and the former, unavailable diagnoses, and to allow for minor inaccuracies in the transcription of the ICD-10 in the CNA. We considered two records linked, when the former four variables were identical in both datasets and the underlying cause of care dependency occurred as a validated diagnosis (see above) in the SHI claims data in 2006 or 2007. Only uniquely linked records were considered successful, that is, only one record from the CNA data was linked to only one record in the SHI claims data. Duplicate linked records were passed on to stage three of the linkage process.

Stage 3 equalled stage 2, except that the underlying cause of care dependency ICD-10 was not truncated. This adds further potential for differentiation between individuals having all other attributes in common. Records not uniquely assigned in this stage were considered not linkable.

### Statistical analyses
Initially, we examined sociodemographic attributes of the two datasets provided. To keep track of the linkage

process, we did interim descriptive analyses using the working datasets. For final assessment of the linkage success, we used summary and descriptive statistics and compared linked with non-linked observations and the base cohort by sex, age group, federal state and care level.

Furthermore, we conducted multivariate logistic regression with these same attributes as independent variables and the linkage success as the dependent variable. We report ORs and 95% CIs for each attribute. Categories with the largest number of observations were used as reference.

For plausibility checking, we examined the chronological order of diagnoses in SHI claims data and the diagnoses stated as the underlying cause of care dependency. Linkage and analyses were conducted using SAS software, V.9.4.[22]

### Patient and public involvement
Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

### RESULTS
The key demographics of both datasets are presented in table 1. The excess 30 000 observations from the CNA data compared with the SHI data may be attributed to the assessments of non-AOK insured individuals in Bremen and Lower Saxony. Besides, counts from each federal state are comparable, except Rhineland-Palatinate presenting twice as many records in the CNA than in the SHI claims data and Hamburg and Schleswig-Holstein the contrary for unknown reasons. Relative distributions among sex, age and care level show only minor deviations between both datasets.

### Linkage process
We found 91 571 CNAs unique by MDK region, sex, date of birth and care level compared with 77 782 records in the SHI claims data also fulfilling these criteria. 50 970 of these could be unambiguously assigned with each other, corresponding to 32.2% of the entire cohort. The remaining records were passed on to the second stage.

Adding the underlying cause of care dependency truncated to 3-digit ICD-10 to the key variables region, sex, date of birth and care level, we found 97 957 assessments uniquely identifiable in the remaining CNA dataset. Of these, 68 625 assessments could be linked to 107 099 individuals in the AOK dataset. However, only 13 816 matches were distinct with a one-to-one relationship between both respective observations in both datasets. The remaining resulted in one-to-many or many-to-many matches, thus not distinguishable between different individuals. Distinct matches are equivalent to 8.7% of the study cohort defined by the SHI data. Where duplicates in the key variables occurred, observations were passed on to the third stage.

**Table 1** Sociodemographic characteristics of the two datasets

| | SHI claims data | | CNA data | |
|---|---|---|---|---|
| | n | % | n | % |
| **Sex** | | | | |
| Male | 57 582 | 36.4 | 67 592 | 35.8 |
| Female | 100 487 | 63.6 | 119 027 | 63.0 |
| Missing | 0 | 0.0 | 2316 | 1.2 |
| **Age** | | | | |
| 65–69 | 14 423 | 9.1 | 16 794 | 8.9 |
| 70–74 | 21 905 | 13.9 | 25 220 | 13.4 |
| 75–79 | 33 472 | 21.2 | 38 951 | 20.6 |
| 80–84 | 40 587 | 25.7 | 48 800 | 25.8 |
| 85–89 | 30 097 | 19.0 | 36 920 | 19.5 |
| 90–94 | 13 973 | 8.8 | 17 680 | 9.4 |
| 95–99 | 3411 | 2.2 | 4304 | 2.3 |
| 100–104 | 196 | 0.1 | 260 | 0.1 |
| 105–109 | 5 | 0.0 | 6 | 0.0 |
| **Region** | | | | |
| Berlin & Brandenburg | 16 060 | 10.2 | 16 275 | 8.6 |
| Baden-Wuerttemberg | 19 617 | 12.4 | 16 763 | 8.9 |
| Bavaria | 22 410 | 14.2 | 20 117 | 10.7 |
| Bremen | 1548 | 1.0 | 3669 | 1.9 |
| Hesse | 12 771 | 8.1 | 12 079 | 6.4 |
| Hamburg & Schleswig-Holstein | 6838 | 4.3 | 2841 | 1.5 |
| Mecklenburg-West Pomerania | 6838 | 4.3 | 6702 | 3.6 |
| Lower Saxony | 1395 | 0.9 | 30 664 | 16.2 |
| North Rhine-Westphalia | 25 089 | 15.9 | 26 444 | 14.0 |
| Rhineland-Palatinate | 8419 | 5.3 | 16 954 | 9.0 |
| Saarland | 2160 | 1.4 | 2273 | 1.2 |
| Saxony | 16 259 | 10.3 | 16 319 | 8.6 |
| Saxony-Anhalt | 9067 | 5.7 | 8989 | 4.8 |
| Thuringia | 9598 | 6.1 | 8846 | 4.7 |
| **Level of care** | | | | |
| I | 108 649 | 68.7 | 131 930 | 69.8 |
| II | 40 581 | 25.7 | 46 920 | 24.8 |
| III | 8839 | 5.6 | 10 085 | 5.3 |
| Total | 158 069 | 100.0 | 188 935 | 100.0 |

CNA, Care Needs Assessment; SHI, Statutory Health Insurance.

In the third stage, a procedure similar to the second stage was applied but with the non-truncated ICD-10 of the underlying cause of care dependency. Of 85 011 assessments in the CNA data, 13 146 could be linked to 93 283 remaining individuals in the AOK dataset. Thereof, 1315 were one-to-one matches, thus in the third stage, 0.8% of the study cohort could additionally be successfully matched with assessments. Overall, 41.8% (n=66 101) of the study cohort could be matched with CNA data.

## Evaluation of linkage results

The proportions of AOK insured of the study cohort with successfully linked CNAs are presented in table 2. For both states with missing day and month of birth, Bavaria and Baden-Wuerttemberg, the proportion of observations linked was near zero. A noticeably low share can also be found for the region of Hamburg and Schleswig-Holstein with 20.2%. The highest proportions could be found in Saarland (74.4 %) and Bremen (73.6 %), two of the smallest federal states of Germany. The percentage

**Table 2** Successfully linked records by stage of linkage process and region

| Region | No linkage (%) | Linkage successful | | | |
| --- | --- | --- | --- | --- | --- |
| | | All stages (%) | Stage 1 (%) | Stage 2 (%) | Stage 3 (%) |
| Berlin & Brandenburg | 43.1 | 56.9 | 41.2 | 14.3 | 1.4 |
| Baden-Wuerttemberg | 99.9 | 0.1 | 0.0 | 0.0 | 0.0 |
| Bavaria | 99.3 | 0.7 | 0.0 | 0.6 | 0.1 |
| Bremen | 26.4 | 73.6 | 69.4 | 4.1 | 0.2 |
| Hesse | 41.5 | 58.5 | 47.6 | 10.2 | 0.7 |
| Hamburg & Schleswig-Holstein | 79.8 | 20.2 | 17.9 | 2.3 | 0.1 |
| Mecklenburg-West Pomerania | 34.1 | 65.9 | 58.0 | 7.4 | 0.5 |
| Lower Saxony | 53.0 | 47.0 | 33.1 | 12.7 | 1.2 |
| North Rhine-Westphalia | 47.0 | 53.0 | 34.5 | 16.8 | 1.8 |
| Rhineland-Palatinate | 31.0 | 69.0 | 53.2 | 14.6 | 1.2 |
| Saarland | 25.6 | 74.4 | 71.4 | 2.8 | 0.1 |
| Saxony | 42.3 | 57.7 | 43.7 | 12.6 | 1.4 |
| Saxony-Anhalt | 36.2 | 63.8 | 54.7 | 8.7 | 0.4 |
| Thuringia | 40.0 | 60.0 | 50.1 | 8.8 | 1.1 |
| **Total** | 58.2 | 41.8 | 32.2 | 8.7 | 0.8 |

Data: study cohort defined through SHI claims data (n=158 069).
SHI, Statutory Health Insurance.

linked in the second stage was higher in states having a larger population with up to 16.8% in North Rhine Westphalia. Additional matches in the third stage were under 2% for each region.

Looking at the differentials in linkage success by sex, we see a lower overall rate in women, comprising nearly two thirds of the cohort, than in men (table 3). Adding the underlying cause of care dependency in the second and third stage resulted in more matches in women than men, although not making up for the overall higher linkage rate for men.

Age-specific linkage rates were higher on the upper and lower age boundaries with 51.5% for 66 year olds and 100.0% for 108 year olds (see figure 2). The lowest rate could be observed at the age of 85 with 33.8 %, which also is the age with the highest frequency in the cohort. Figure 2 shows the inversely proportional relationship between linkage success and number of individuals

present in the data based on their age. The probability of linkage shows to be higher, when there are less people of the same age and vice versa. Pearson's correlation coefficient for this relationship is −0.67.

Figure 3 presents the resulting composition of the linked dataset compared with the study cohort based on SHI claims data. The relative frequency of some categories varies between both datasets due to the aforementioned differential linkage rates. Large distortions in the linked dataset are not apparent.

For the purpose of validation and plausibility checking, we further examined the relationship between the underlying cause of care dependency coded as an ICD-10

**Table 3** Successfully linked records by stage of linkage process and sex

| Sex | No linkage (%) | Linkage successful | | | |
| --- | --- | --- | --- | --- | --- |
| | | All stages (%) | Stage 1 (%) | Stage 2 (%) | Stage 3 (%) |
| Male | 53.0 | 47.0 | 39.2 | 7.2 | 0.6 |
| Female | 60.9 | 39.1 | 28.3 | 9.7 | 1.1 |
| All | 58.0 | 42.0 | 32.2 | 8.8 | 0.9 |

Data: study cohort defined through SHI claims data (n=158 069).
SHI, Statutory Health Insurance.
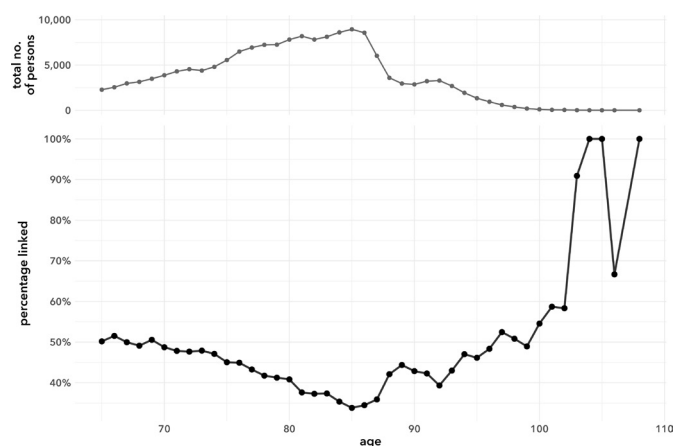


**Figure 2** Percentage of successfully linked individuals and total number of individuals in the dataset by age. Data: study cohort defined through SHI claims data (n=158 069). SHI, Statutory Health Insurance.
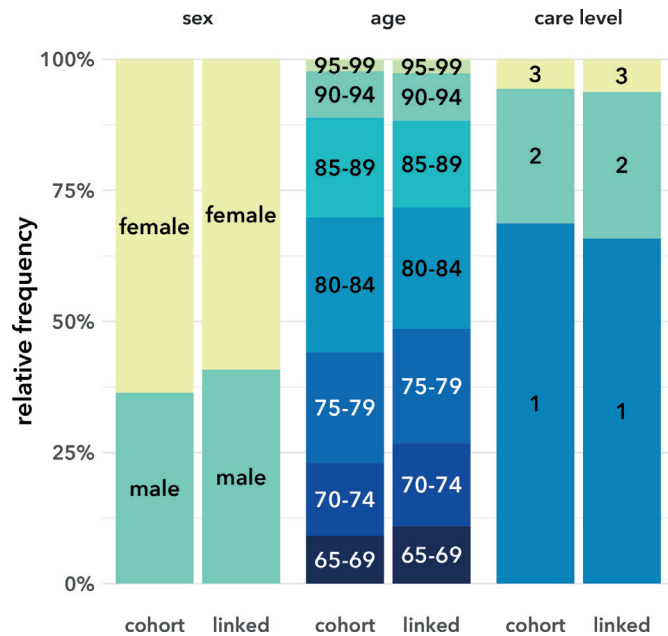
**Figure 3** Comparison of sociodemographic composition of study cohort and resulting linked dataset data: Cohort: study cohort defined through SHI claims data (n=158 069); linked: study cohort with successfully linked care needs assessments (n=66 101). SHI, Statutory Health Insurance.

diagnosis by the MDK in the CNA and the inpatient and outpatient diagnosis from SHI claims data. For those individuals linked in the first stage (without consideration of the underlying cause of care dependency), we reviewed the proportion of valid diagnosis from CNA data that could also be found in the SHI claims data. We found 57.0% of individuals with a valid ICD-10 diagnosis in CNA data had at least one occurrence of the same diagnosis in in- and outpatient SHI claims. The respective diagnosis was coded for 56.3% at least once and for 43.2% at least twice within four quarters preceding or following the date of CNA. First time diagnoses occurred mostly prior to the assessment, with 66.4% of linked individuals from the first stage with valid diagnosis. For an additional 25.4%, it occurred in the quarter of the CNA.

We repeated this analysis for individuals linked in the second and third stage and found that 61.1% of individuals having the diagnosis coded up to four quarters before the assessment for the first time and 26.4% in the same quarter as the assessment took place.

## Determinants for successful linkage

To assess individual factors contributing towards successful linkage a multivariate logistic regression was conducted with sex, level of care, region of residence and age group as the independent variables. The results are presented in table 4. We found that men had a 1.5-fold higher chance for successful linkage than women. Compared with care level 1, individuals with care level 2 had a mildly higher chance (OR=1.25) of linkage with CNA data, while in care level 3 chance was slightly reduced (OR=0.92). Considering the regions, we find ORs of about zero for

**Table 4** Estimates from logistic regression with linkage success as dependent variable.

| Independent variables | OR (95% CI) |
|---|---|
| Sex | |
| Female | Ref. |
| Male | 1.50 (1.46 to 1.54) |
| Care level | |
| 1 | Ref. |
| 2 | 1.25 (1.21 to 1.28) |
| 3 | 0.92 (0.88 to 0.97) |
| Region | |
| North Rhine-Westphalia | |
| Berlin | 1.21 (1.16 to 1.26) |
| Baden-Wuerttemberg | 0.00 (0.00 to 0.00) |
| Bavaria | 0.01 (0.01 to 0.01) |
| Bremen | 2.60 (2.31 to 2.93) |
| Hesse | 1.28 (1.23 to 1.34) |
| Hamburg | 0.21 (0.2 to 0.23) |
| Mecklenburg-West Pomerania | 1.75 (1.65 to 1.85) |
| Lower Saxony | 0.78 (0.7 to 0.87) |
| Rhineland-Palatinate | 1.98 (1.88 to 2.09) |
| Saarland | 2.63 (2.38 to 2.91) |
| Saxony | 1.25 (1.2 to 1.31) |
| Saxony-Anhalt | 1.63 (1.55 to 1.71) |
| Thuringia | 1.39 (1.33 to 1.46) |
| Age group | |
| 80–84 | Ref. |
| 65–69 | 1.83 (1.75 to 1.92) |
| 70–74 | 1.57 (1.51 to 1.64) |
| 75–79 | 1.26 (1.22 to 1.3) |
| 85–89 | 0.98 (0.95 to 1.02) |
| 90–94 | 1.46 (1.39 to 1.53) |
| 95–99 | 2.21 (2.02 to 2.41) |
| 100–104 | 5.43 (3.47 to 8.49) |
| 105–109 | 85.07 (0.13 to Inf.) |

Data: study cohort defined through SHI claims data (n=158 069). Ref., reference category; SHI, Statutory Health Insurance.

Bavaria and Baden-Wuerttemberg, where date of birth was missing in the data and little observations could be linked due to a lack of uniqueness, compared with the region North Rhine-Westphalia. Hamburg also had a considerably reduced OR of 0.21. Individuals from Lower Saxony had a moderately reduced chance of successful linkage (OR=0.78). The chance of linkage notably elevated for the two states Bremen (OR=2.60) and Saarland (OR=2.63). The remainder of 7 regions, the OR was between 1 and 2, each compared with North Rhine-Westphalia. With regards to age groups, we found that nearly all age groups had a higher chance of successful

linkage compared with the age group of 80–84 year olds, the group with the largest number of cases. The only exception is made by the group of 85–89 year olds, where no statistically significant difference compared with the reference group is present. Especially in the older age groups of 95 years and above, which have considerably lower numbers of insured, the OR is elevated with an up to 5.43-fold chance of linkage for 100–104 years. For the oldest age group of 105 and above, no statistically significant difference could be found.

## DISCUSSION

In this article, we describe the conduction of a deterministic record linkage of German SHI claims data with CNA data. Lacking a universal and unique personal identifier available in both datasets, we were restricted to the four key variables region, sex, date of birth and care level. Additionally, we could add the underlying cause of care dependency, coded as ICD-10-GM, from the CNA data, which is supposed to correspond to the diagnoses present in the SHI claims data. After a stepwise linkage process, we could realise a moderate degree of matches between both datasets between 20.2% and 74.4%, varying between federal states. Therefore, potential for improvement is evident, as we expect both included original datasets to be complete and contain the identical individuals after applying study inclusion and exclusion criteria. Deviations between the datasets may only be reasoned by erroneous or missing data, or a lack of documentation or understanding for the data resulting in an incorrect application of criteria. In this study the former was present in Bavaria and Baden-Wuerttemberg, where day and month of birth were missing, the latter was the case for failing to select only AOK insured for Bremen and Lower Saxony from the MDS dataset, where this differentiation could not be made as identifiers for SHI funds did not appear to match the documented scheme. Sociodemographic composition of both datasets was largely comparable and supported the suitability of the data for data linkage, although representativity of the resulting data must be critically reflected for further usage.

As the evaluation of the linkage results indicates, success rate was especially limited, as the available key variables did not suffice to distinguish between all individuals in the dataset. This was seen in a lower linkage probability, when more individuals shared the same attribute. To our knowledge, there is only one study in Germany having performed data linkage with comparable data.[23] This study only used data from the federal state of Hesse and achieved a proportion of 75.8% successfully linked CNA. Compared with 58.5% people in Hesse we could link to CNA data, the higher amount may be attributed to the additional linkage variables postcode and data of assessment, the authors were able to use, while they only included age, rather than the entire date of birth as in this study. Considering an expected increase in the number of care dependent people due to demographic change and LTCI reforms,[7] the yield of the presented exact method may result in lower linkage rates in future studies, as the ability to distinguish between individuals is lower when more people share the same limited set of attributes.

With the record linkage affected by sociodemographic attributes, this has an impact on the resulting combined dataset, which may not be representative for the original data regarding these and further attributes, contributing to a successful linkage. In our dataset, we do find an under-representation of women, and those around 85 years old. However, this only led to a slight distortion in the frequency distribution. The largest absolute change was under five percentage points with 63.6 % women in the original and 59.2% in the linked data. In multivariate analyses these distortions show more pronounced, as the individual's chance of successful record linkage is significantly determined by their sociodemographic attributes. As distinct attributes enhance the chance of a successful linkage, comparable effects on the representativity are to be expected for the underlying cause of care dependency, with rare diseases being potentially overrepresented and more common diseases (eg, dementia) being under-represented.

Considering these results, representativity of linked datasets must be reflected and incorporated into further analyses. The primary outcome of the underlying project is defined as the institutionalisation of care dependent people. Subsequent analyses[10–13] focus on determinants from health care and the individuals' care arrangement on the risk of nursing home admission after first onset of care dependency. A selection introduced by record linkage may impair the validity of these analyses. Here, we saw non-matches in the second and third stage being largely attributable to a lack of distinction with the available linkage variables. The existence of two individuals with similar sociodemographic attributes alone should not be associated with the probability of institutionalisation. However, a resulting higher proportion of men than women ending up in the final dataset may have influence on the external validity of subsequent analyses on associations with nursing home admission. Therefore, it is indicated to perform these analyses stratified by known determinants of successful record linkage or include respective variables as covariates in inferential statistics. With nearly no observations from Bavaria and Baden-Wuerttemberg having been matched, these states need to be excluded in further analyses and affect the representativity of our data for the whole of Germany. This has to be underlined further, as differences in the sociodemographics of health insurance funds are also well known.[24–26]

A common approach for enhancing the proportion of linkable records is the probabilistic record linkage, which does not only depend on the presence of a singular combination of features but takes into account whether two observations have a higher probability of belonging to the same individual than others.[27] The

underlying cause of care dependency poses as the most promising opportunity for this study. Based on the frequency of the coding of the diagnosis in SHI claims data and the assumption, that it should occur before or concurrent with the CNA—as determining a medical diagnosis is not the focus of the assessment—there should be a higher probability for a match, if these criteria are fulfilled. We neglected this approach for the presented study, as we do not have SHI claims data available for all individuals for the time before the CNA. Especially the fact that for only 12.5% of individuals from second or third stage the diagnosis was coded for the first time after the CNA underlines this, as the expected workflow in CNA would be a transfer from existing medical documentation. Additionally, with only 57.0% of included individuals having the ICD-10 diagnosis of the assessed cause of care dependency also present in their SHI data, the validity of these diagnoses should be examined prior to further usage. Prior research showed impaired coding quality of specific diagnoses for various reasons, for example when they do not have an impact on reimbursement.[21]

Based on the results presented we consider the data linkage valid with the aforementioned limitations for regional representativity and methodological caveats. Further validation could comprise examining diagnosis from SHI claims data and the functional status surveyed in CNA, for example, an individual with tetraplegia may be expected no to be mobile on their own, without aids or personal support. However, validation of this kind is only suitable for a very small share of individuals and thus, multiple of these plausibility checks need to be identified.

## CONCLUSION

While less than half of all observations from both datasets—expected to contain the same individuals—could be linked together, we consider this a good result due to the lack of unique or highly distinct identifiers common to both datasets. Without a unique identifier, the specific process of data linkage is largely determined by the available data. Therefore, knowledge about the origins of data is necessary when working with secondary data, especially when performing record linkage, where variables and attributes are only seemingly identical with data emerging from different sources by different processes for different purposes. The presence of an overarching unique identifier in data from the German social system, suitable and available for scientific purposes, could contribute to a better data quality and more efficient research.

**ORCID iDs**
Dominik Domhoff http://orcid.org/0000-0001-5876-0334
Kathrin Seibert http://orcid.org/0000-0002-0422-3637
Susanne Stiefler http://orcid.org/0000-0002-8048-8207
Karin Wolf-Ostermann http://orcid.org/0000-0001-8513-3125

## REFERENCES

1 Lehnert T, Heuchert M, Hussain K. *Stated preferences for long-term care: a literature review. Ageing and society*. Cambridge University Press, 2018: 1–41.
2 Heuchert M, König H-H, Lehnert T. [The Role of Preferences in the German Long-Term Care Insurance - Results from Expert Interviews]. *Gesundheitswesen* 2017;79:1052–7.
3 Santos-Eggimann B, Meylan L. Older citizens' opinions on long-term care options: a vignette survey. *J Am Med Dir Assoc* 2017;18:326–34.
4 Hajek A, Lehnert T, Wegener A, *et al*. [Long-Term Care Preferences Among Individuals of Advanced Age in Germany: Results of a Population-Based Study]. *Gesundheitswesen* 2018;80:685–92.
5 Luppa M, Luck T, Weyerer S, *et al*. Prediction of institutionalization in the elderly. A systematic review. *Age Ageing* 2010;39:31–8.
6 Blümel M, Spranger A, Achstetter K, *et al*. Germany: health system review. *Health Systems in Transition* 2020;22:i–272.
7 Nadash P, Doty P, von Schwanenflügel M. The German long-term care insurance program: evolution and recent developments. *Gerontologist* 2018;58:588–97.
8 World Health Organization. Regional Office for E, European Observatory on Health S, Policies. *Germany: health system review*. Copenhagen: World Health Organization. Regional Office for Europe, 2014: 296.
9 Rothgang H. Social insurance for long-term care: an evaluation of the German model. *Soc Policy Adm* 2010;44:436–60.
10 Domhoff D, Seibert K, Stiefler S, *et al*. Differences in nursing home admission between functionally defined populations in Germany and the association with quality of health care. *BMC Health Serv Res* 2021;21:190.
11 Seibert K, Stiefler S, Domhoff D, *et al*. The influence of primary care quality on nursing home admissions in a multimorbid population with and without dementia in Germany: a retrospective cohort study using health insurance claims data. *BMC Geriatr* 2022;22:52.
12 Stiefler S, Seibert K, Domhoff D. Prädiktoren für den Eintritt in ein Pflegeheim bei bestehender Pflegebedürftigkeit – Eine Sekundärdatenanalyse Im Längsschnittdesign. *Gesundheitswesen* 2021;84:139–53.
13 Seibert K, Stiefler S, Domhoff D, *et al*. [Quality of ambulatory medical care in the context of age and care-dependency: Results of a cross-

sectional analysis of German health claims data]. *Z Evid Fortbild Qual Gesundhwes* 2020;155:17–28.

14 March S, Antoni M, Kieschke J, *et al*. Quo vadis Datenlinkage in Deutschland? Eine erste Bestandsaufnahme. *Gesundheitswesen* 2018;57:e20–31.

15 Gilbert R, Lafferty R, Hagger-Johnson G, *et al*. Guild: guidance for information about linking data sets†. *J Public Health* 2018;40:191–8.

16 von Elm E, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.

17 Czwikla J, Domhoff D, Giersiepen K. [ICD coding quality for outpatient cancer diagnoses in SHI claims data]. *Z Evid Fortbild Qual Gesundhwes* 2016;118-119:48–55.

18 Stausberg J. [Quality of coding in acute inpatient care]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2007;50:1039–46.

19 Erler A, Beyer M, Muth C, *et al*. Garbage in – garbage out? Validität von Abrechnungsdiagnosen in hausärztlichen Praxen. *Gesundheitswesen* 2009;71:823–31.

20 Schubert I, Ihle P, Köster I. Interne Validierung von Diagnosen in GKV-Routinedaten: Konzeption MIT Beispielen und Falldefinition. *Gesundheitswesen* 2010;72:316–22.

21 Wockenfuss R, Frese T, Herrmann K, *et al*. Three- and four-digit ICD-10 is not a reliable classification system in primary care. *Scand J Prim Health Care* 2009;27:131–6.

22 SAS Version 9.4 [program]. Cary, NC, USA 2016.

23 Küpper-Nybelen J, Ihle P, Deetjen W, *et al*. Empfehlung rehabilitativer Maßnahmen Im Rahmen Der Pflegebegutachtung und Umsetzung in Der ambulanten Versorgung. *Zeitschrift für Gerontologie und Geriatrie* 2006;39:100–8.

24 Hoffmann F, Icks A. Unterschiede in Der Versichertenstruktur von Krankenkassen und deren Auswirkungen für die Versorgungsforschung: Ergebnisse des Bertelsmann-Gesundheitsmonitors. *Gesundheitswesen* 2012;74:291–7.

25 Jaunzeme J, Eberhard S, Geyer S. Wie „repräsentativ" sind GKV-Daten? *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2013;56:447–54.

26 Epping J, Geyer S, Eberhard S. Völlig unterschiedlich Oder doch recht ähnlich? die soziodemografische Struktur Der AOK Niedersachsen Im Vergleich Zur niedersächsischen und bundesweiten Allgemein- und Erwerbsbevölkerung. *Gesundheitswesen* 2021;83:S77–86.

27 Sayers A, Ben-Shlomo Y, Blom AW, *et al*. Probabilistic record linkage. *Int J Epidemiol* 2016;45:954–64.