Assessing the re-identification potential of health care data for people with statutory health insurance in Germany (preliminary results)

Hannah Lobert * and Adam Giza* and Jochen Dreß*

* German Institute for Medical Documentation and Information / Deutsches Institut für Medizinische Dokumentation und Information (DIMDI),
  [jochen.dress][hannah.lobert]@dimdi.de

Abstract: The assessment of re-identification risk is a crucial task for any organisation that provides access to personal data for research purposes. Particularly, the sensitive nature of health care data obliges institutions dealing with them to ensure confidentiality and prevent re-identification of individuals.

There are several methods to measure re-identification risk or – to be more precise – re-identification potential (Duncan et al. 2011). We speak here of re-identification potential rather than re-identification risk, because the methods do not take into account all relevant factors (Desai et al. 2016). In this paper, we present the results of applying the uniqueness measure to German statutory health insurance data. The data comprise records originating from about 70 million people with statutory health insurance in Germany. They contain socio-demographic information, information on outpatient medication, inpatient and outpatient diagnoses and health insurance expenses. We investigated attribute patterns formed by sex, year of birth, and diagnoses coded according to the ICD-10-GM with their respective qualification (suspected, confirmed, condition after, exclusion, other). We considered four scenarios with different levels of detail regarding the diagnoses. For the most detailed scenario 1, patterns were formed by sex, year of birth, and one diagnosis with its qualification per quarter for each quarter of the reporting year in chronological order (four diagnoses in total). For the least detailed scenario 4, patterns were formed by two diagnoses documented in different quarters of the reporting year without information on their chronological order.

We found that in general there were many attribute patterns, and the number of patterns was higher for women and for older individuals. In addition, the proportion of unique or rare patterns was very high. In the most detailed scenario 1, almost the entire dataset consisted of unique patterns (e.g. 95.4 % unique patterns for individuals born in 1960). Although the frequency of patterns was higher for women and older individuals, the proportion of unique patterns was slightly higher for men and younger individuals (women born in 1960 had 95.0 % unique patterns, men born in 1960 had 96.1 %, women born in 1980 had 96.9 % and men born in 1980 had 98.7 % unique patterns). In scenario 1, more than half of the people covered by statutory health insurance had 20 or more unique attribute patterns (e.g. 71.0 % for women born in 1960, 64.8 % for 1970 and 64.5 % for 1980).

The calculated re-identification potential is high, even in scenarios incorporating less detailed data. Further investigations should be conducted with other scenarios by modifying the patterns, e.g. taking into account the prescribed and invoiced outpatient drugs and / or using just the three left characters of the ICD code. In future research, the results will be supplemented with record linkage simulations. These will also take into account result sets, i.e. data derived from the data. Based on the results we hope to find a way to better estimate the re-identification risk of the data by incorporating the re-identification potential, the intended utilisation and the settings in which the data and the result sets are distributed and handled.
1 Introduction

The assessment of re-identification risk is a crucial task for any organisation that provides access to personal data for research purposes. Particularly, the sensitive nature of health care data obliges institutions dealing with them to ensure confidentiality and prevent re-identification of individuals. Re-identification risk assessment has been studied by numerous researchers (e.g. Domingo-Ferrer et al. 2015; Elliot et al. 2002) and several studies (e.g. El Emam et al. 2011; El Emam et al. 2013) have examined the special case of the health care sector. In this sector, patient data as personal data are particularly sensitive: Naturally, one does not want the employer, the bank or even the family to know about an illness that might or might not predict a chronic illness or turn into one without having acted upon it oneself. Thus, breaches of confidentiality, be it re-identification or attribution, are taken very seriously.

This paper will present the application of the uniqueness measure to health care data. The uniqueness measure is one way to calculate the re-identification potential. We speak here of re-identification potential rather than re-identification risk, because the method – like others – does not take into account all relevant factors (Desai et al. 2016). In this study, we will demonstrate how easily health care information containing only a few diagnoses, the age and the sex of a person could be misused to pinpoint that person.

In section 1, we will present the dataset and review potentially suitable risk assessment methods for these data. In section 2, we will outline the procedure of analysis in detail, and section 3 deals with the results. Section 4 concludes with the bigger picture and directions for future research.

1.1 The dataset

This paper presents an application of measuring re-identification potential for the special case of German statutory health insurance data (henceforth referred to as DaTraV data). The DaTraV data comprise records originating from about 70 million people with statutory health insurance in Germany. Health insurance is mandatory for every German resident. Most of the people living in Germany have a statutory health insurance; only about 10-12% of the population are members of a private health insurance company (10.6% in 2017; PKV, 2018).

The German statutory health insurance data are generated by the resident physicians, hospitals and pharmacies invoicing their services to the German statutory health insurance funds. The DaTraV data are collected by the German Federal Insurance Office from every German statutory health insurance company, with the purpose of compensating expenses between the companies. To fulfil a tertiary purpose, the data are then transferred to two sub-organisations within the DIMDI. The trust centre receives only the data needed to calculate permanent pseudonyms. The data processing centre receives temporary pseudonymised health care data and joins these data with the
calculated permanent pseudonyms. It offers services to authorised institutions that enable the institutions to analyse the data and obtain the result sets. Therefore, the data processing centre has to ensure confidentiality. Normally, the data are provided in an aggregated manner; however, in well-founded exceptional cases, the pseudonymised data can be analysed on-site. The DaTraV data contain socio-demographic information, information on outpatient medication, inpatient and outpatient diagnoses and health insurance expenses for reporting years 2008-2016.

1.2 Review of potentially suitable risk assessment methods

An extensive body of research has been conducted in the field of risk assessment. Literature on risk assessment classifies methods into measures of individual risk and global file-level risk (Duncan et al. 2011). To gain insights into how to select a suitable method for measuring re-identification risk with the categorical DaTraV data, we conducted a survey with experts in the field of statistical confidentiality. In this survey, we discussed the uniqueness measure, the special uniqueness detection algorithm (SUDA), measuring file-level risk based on individual risk, loglinear models, record linkage and an information-theoretic approach. In the following, we present the results of this expert survey together with our own reflections.

Most of the work on risk assessment for microdata focused on either the concept of uniqueness or on record linkage simulation studies. The uniqueness measure is defined as follows: “A record is unique on a set of key variables if no other record shares its values for those variables.” (Duncan et al. 2011, p. 42). There is a distinction between population uniqueness (unique in the population) and sample uniqueness (unique in the sample). This distinction leads to related methods for risk assessment, like the proportion of sample uniqueness that are population unique, the probability of a correct match given a unique match developed by Skinner and Elliot 2002, or the detection of “special uniqueness”, which are also unique on a subset of variables and can be measured by the special uniqueness detection algorithm SUDA (Elliot et al. 2002; Manning et al. 2008). It is also possible to infer the frequency of attribute combinations in the population from frequencies in the sample to estimate individual re-identification risk (Skinner, Holmes 1998; Benedetti et al. 2003).

Dankar et al. 2012 give an overview of metrics derived from uniqueness as well as estimators of population uniqueness if sample uniqueness is given. A commonly used measure for population uniqueness is simply the proportion of records that are unique in the population. Uniqueness and population uniqueness in particular can serve as a proxy for re-identification risk of an ensemble of records (Duncan et al. 2011). It is thus useful for determining the overall risk of cases in a file. At file level, the risk potential can be determined by the percentage of either unique or very rare attribute patterns. Nevertheless, it can also be useful for assessing individual risk: If the attribute combination of a record is unique or the specific combination is very rare, this record is at risk of re-identification. The uniqueness measure has several advantages: Firstly, it provides the basis for various risk measures. Secondly, its informative value is great and
intuitive. Thirdly, it can easily be implemented even with the complex pattern structure of the DaTraV data. That is why we consider the uniqueness measure in our present paper.

As mentioned before, the algorithm SUDA can identify special uniques (Elliot et al. 2006). The first step of the algorithm is the generation of the power set of the attributes in the dataset. After that, records are grouped by the different attribute combinations in the set. Every record is then scored according to the number of unique attribute combinations of the power set. The SUDA algorithm as well as its successor SUDA2 is computationally intensive, although there have been optimisations to speed up the computation (Templ et al. 2015). Another issue specific to the DaTraV data is that one record can have more than one attribute pattern, because one person can have more than one diagnosis (explained in more detail below). Implementing this with SUDA would complicate the procedure further and prolong computation time.

A different approach for estimating re-identification risk is actively attempting to link records from two different files and then determining the proportion of correct matches. This can either be achieved with an external dataset or with the same dataset before and after disclosure control methods have been applied (Duncan et al. 2011). There are distance-based or probabilistic record linkage techniques (Domingo-Ferrer, Torra 2002). In distance-based record linkage, pairs of records with a minimum distance measure are matched. In probabilistic record linkage, the ratio between conditional probabilities is calculated: The probability of having identical values on the key variables is conditioned on the assumptions of the record pair being either a match or not. Record linkage is also suitable for incorporating and testing different assumptions about the attacker’s knowledge and strategy (e.g. see Domingo-Ferrer et al. 2015 for maximum-knowledge attacks). Thus, record linkage is regarded as a sensible complement for measuring re-identification risk with the DaTraV data.

In addition to these methods, risk can also be assessed using loglinear models (Skinner, Holmes 1998). This method can be used to estimate the number of sample uniques that are also population unique. Therefore, the expected cell counts of the corresponding contingency table are modelled by means of a loglinear model. The parameters of the model are estimated by maximum likelihood estimation (Skinner, Holmes 1998; Ichim, Foschi 2011). One advantage of this method is the possibility to incorporate model assumptions in the estimation. But since the sample of the DaTraV data covers almost the entire population, the estimation of such a model would be of no additional value. Furthermore, the method cannot estimate the risk of individual records. Thus, loglinear models are considered unsuitable for measuring re-identification risk for every record in the DaTraV file. Nevertheless, they could be a valuable approach if samples of the data were to be investigated.

Another method to estimate re-identification risk is an information-theoretic approach (Antal et al. 2014; Shlomo et al. 2015). This measure is only suitable for tabular data and not for microdata, and it measures attribute disclosure rather than identity
disclosure. The information-theoretic approach is based on entropy and conditional entropy. Disclosure risk can be estimated by the reciprocal of an entropy measure that represents the information loss between the data processor’s information and the information contained in the dataset at risk. After the expert survey had been conducted, we decided to measure individual re-identification risk in microdata, which is why this approach was not considered for our present research. For the same reason, the combination of individual risk measures to assess file-level risk has not been pursued.

The expert survey helped with deciding on a method for assessing the re-identification potential of the DaTraV data. The best applicable methods were identified to be the uniqueness measure as well as record linkage simulation. These methods are to be implemented on the DaTraV data. In this paper, we describe the methodology and the preliminary results obtained by the uniqueness measure. In a future paper, we aim to compare these results to the findings of a record linkage simulation study.

2 Methods

For the paper at hand, the assessment of re-identification potential has been conducted using the uniqueness measure. In this section, we will describe the application to the DaTraV data, the selection of the attributes used to construct the patterns (key variables) and the definition of risk scenarios.

2.1 Application of the uniqueness measure to the DaTraV data

The DaTraV data represent the German population with statutory health insurance to a great extent. Due to this fact, the difference between the population and the sample is almost negligible. Hence, we assume equivalence of population uniqueness and sample uniqueness for the DaTraV data. If a record is found to be sample unique, we therefore assume it is population unique as well.

This poses further complications for data confidentiality and for the investigation of uniqueness: Population uniqueness is a more limited concept than sample uniqueness. If a record is found to be sample unique, there is still some possibility that it is not unique in the population. However, if the sample almost corresponds to the entire population, a record that is sample unique is at high risk of being re-identifiable or to allow re-identification of people when data are linked.

2.2 Selection of attributes used to construct the patterns (key variables)

The first step for assessing uniqueness in the present study was the selection of key variables from the DaTraV data. Knowledge of key variables facilitates re-identification in that these variables can be used to match records from two datasets (Duncan et al. 2011). The choice of key variables depends on the definition of risk scenarios (see below) that describe specific knowledge that could lead to confidentiality breaches. Key
variables are often of a socio-demographic nature or contain very specific personal information.

The DaTraV data comprise only a few socio-demographic attributes, namely sex and year of birth for all reporting years. Additionally, for reporting years 2009 and 2010, geographic information is available with the municipality code. Moreover, the DaTraV data contain very specific personal information in the form of documented diagnoses. For each diagnosis, the data comprise the documented ICD code (DIMDI 2018), the quarter of documentation, the qualification or type of diagnosis and the sector in which the diagnosis was documented.

The selection of key variables for the present study resulted from a careful consideration of the knowledge someone might have and which could lead to a confidentiality breach. The year of birth and the sex of a person are frequently listed in registries or other databases, and these attributes can easily be known by other persons. The specific diagnosis as well as the corresponding circumstances of diagnostics is also something one person might know about another person. This can either be the case if that person has access to data, e.g. at a hospital or at a health insurance company, or if that person knows the other person well. These considerations led to the selection of the following key variables:

1. Year of birth (numeric with four digits)
2. Sex (binary female/male)
3. Documented ICD codes (alphanumeric with 1 character and 2-4 digits, e.g. E11 for diabetes mellitus type 2)
4. Reporting year (numeric with four digits)
5. Quarter of diagnosis (numeric 1-4)
6. Qualification (string suspected/confirmed/condition after/exclusion/other)
7. Type of diagnosis (binary primary/secondary)
8. Sector (binary outpatient/inpatient)

2.3 Definition of risk scenarios

The assessment of disclosure risk is always dependent on context factors. It is not possible to estimate the risk of re-identification without considering the background, the knowledge, or the goals of a person that might process the data in a way that could lead to a confidentiality breach. That is why a proper definition of risk scenarios is so important. A risk scenario should describe a realistic situation that can pose a threat to confidentiality. It can then be used to assess the disclosure risk as well as the effectiveness of disclosure control methods in that specific situation. Duncan et al. (2011) list several properties of a potential attack on or misuse of the data that should be taken into account when defining risk scenarios and key variables.
The risk scenarios described as follows all consider the key variables as stated above, but vary in the level of detail a person knows about these variables. For our study, we defined four risk scenarios for outpatient diagnoses during one reporting year. For every risk scenario, we assumed that year of birth and sex are known. Additionally, the corresponding qualification for each diagnosis was assumed to be known.

**Scenario 1:** For each quarter of one year, there is one known diagnosis. For example, a data processor knows that in the first quarter, a person was diagnosed with diabetes, in the second and third quarter that person had a bladder infection and in the fourth quarter an upper respiratory infection.

**Scenario 2:** For two quarters of one year, there is one known diagnosis per quarter. For example, a data processor knows that in the first quarter, a person was diagnosed with diabetes, and in the second quarter with a bladder infection.

**Scenario 3:** For each half-year of one year, there is one known diagnosis. For example, a data processor knows that in the first half-year, a person was diagnosed with diabetes, and in the second half-year with a bladder infection.

**Scenario 4:** For two quarters of one year, there is one known diagnosis per quarter. For example, a data processor knows that in one quarter, a person was diagnosed with diabetes, and in another quarter with a bladder infection. However, the chronological order of the diagnoses is unknown.

2.4 Procedure

The investigation of uniqueness requires the generation and counting of attribute patterns. The intuitive approach would have been to form the patterns with all diagnoses of one reporting year. However, it is not realistic to assume a data processor would know each specific diagnosis together with the respective quarter and qualification. That is why the risk scenarios cover four diagnoses at most. The definition of the scenarios implies that each record in the database can have more than one attribute pattern. While most risk assessment methods cannot account for this, uniqueness can still be determined.

The DaTraV data contain almost 3 billion records for each reporting year. Analysing all of these data in one single batch is neither computationally efficient nor sensible regarding the risk scenarios described above. Thus, the present analysis considered only data from one (longer past) reporting year and a selection of years of birth. The preliminary results are presented for birth cohorts 1960, 1970 and 1980. The data were examined in batches separated by sex and the chosen years of birth.

3 Results

The first aim of the analysis was to find the number of attribute patterns in the data. Table 1 shows the number of records in the dataset and the respective number of
attribute patterns for women and men born in 1960, 1970, and 1980. Women have more attribute patterns than men, and the number of patterns increases with age.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of insured women with at least one documented diagnosis</td>
<td>Number of attribute patterns</td>
</tr>
<tr>
<td>born 1980</td>
<td>428,979</td>
<td>3,015,917,359</td>
</tr>
<tr>
<td>born 1960</td>
<td>544,015</td>
<td>19,050,562,519</td>
</tr>
</tbody>
</table>

Tab. 1: Number of attribute patterns in scenario 1 for different cohorts in the DaTraV dataset for one reporting year.

The higher number of patterns for women reflects that more women than men had at least one documented diagnosis in the reporting year (35,081,929 women compared to 29,019,186 men) and that women in general had more diagnoses than men in the reporting year (609,065,317 diagnoses for women compared to 369,028,036 for men). The increase of attribute patterns with age is explicable as well: Older people tend to have more diagnoses and thus more diagnosis combinations and attribute patterns.

Bearing the number of attribute patterns in mind, we looked at how often each pattern occurred in the data. Figure 1 shows the proportion of patterns which occur exactly once, twice, ..., ten times for the cohort born in 1960. It can be seen that for scenario 1, the proportion of unique attribute patterns is close to 100 % (95.4 %), whereas for the other three scenarios, it is around 60 % (60.4 % for scenario 2, 60.7 % for scenario 3, and 57.8 % for scenario 4). The proportion drops sharply for patterns that occur twice in the dataset. For scenario 1, it is already close to 0 % (3.0 %). The other three scenarios are again more alike: The proportion of patterns that occur twice is about 14 % (14.1 % for scenario 2, 14.0 % for scenario 3, and 14.3 % for scenario 4). For patterns that occur three or more times, the number of patterns decreases slowly to almost 0 %.
Fig. 1: Proportion of attribute patterns depending on the frequency of occurrence in the reporting year for persons with year of birth 1960.

The incidence of an attribute pattern depends on the scenario: The proportion of unique patterns is highest in scenario 1. In the other scenarios, the proportion of patterns that occur more than once is substantial and decreases with increasing frequency of the pattern. Since scenario 1 considers combinations of four specific diagnoses, their qualification and their corresponding quarters, one can imagine having mostly unique attribute patterns. Intuitively, it is highly unlikely that e.g. two women born in 1960 would have received exactly the same diagnoses in each quarter of the reporting year.

If we compare men and women with the same year of birth, there is no considerable difference between the graphs produced by the different scenarios (fig. 2). This is also true for other years of birth. Thus, sex seems to have an impact on the general number of attribute patterns, but not necessarily on the number of unique attribute patterns or patterns that are generally rare in the dataset.
Fig. 2: Proportion of attribute patterns depending on the frequency of occurrence in the reporting year for women and men with year of birth 1960.

Comparing the exact proportion of unique attribute patterns across sex and years of birth, we can see that the proportion of unique patterns is slightly higher for men than for women and that the proportion of unique patterns decreases slightly with age (see fig. 3). This appears to be the reverse finding than in table 1. However, table 1 emphasizes that women and elderly people in general have more attribute patterns, but they are not always unique. For instance, the pattern “Q1: bladder infection”, “Q2: nothing”, “Q3: nothing”, “Q4: nothing” might well apply to more than one of the almost 430,000 women born in 1980.
We also investigated how many of the unique patterns were attributable to a single person. It is perfectly imaginable that although there are many unique attribute patterns, they could belong to only very few persons with many (unusual) diagnoses. As figure 4 shows, this is partly true for scenario 1. More than half of the insureds had 20 or more unique attribute patterns (e.g. 71.0% for women born in 1960, 64.8% for 1970 and 64.5% for 1980). The proportion was calculated with respect to the number of insureds in the birth cohort that had at least one outpatient diagnosis in the reporting year. In the other three scenarios, the proportion of records with many unique attribute patterns is lower. This is also true across birth cohorts, whereby the percentage of people with at least 50 unique attribute patterns decreases with age (more visible for scenario 1). Figure 5 displays that this still holds if women and men are considered separately. Nevertheless, the proportion of women with many different attribute patterns tends to be higher.
Proportion of insured persons with at least X unique attribute patterns for each year of birth

<table>
<thead>
<tr>
<th></th>
<th>1960</th>
<th>1970</th>
<th>1980</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4: Proportion of insured persons with at least X unique attribute patterns in the reporting year for birth cohorts 1960, 1970 and 1980.
4 Discussion

4.1 Summary of the findings

In our paper, we investigated the uniqueness of attribute patterns for health care data of people with statutory health insurance in Germany. Our aim was to get an idea about the re-identification potential of the data we make available for research and controlling. The attribute patterns we investigated were formed by sex, year of birth, diagnoses coded according to the ICD-10 GM, their qualification and when they were made. We considered four scenarios with a different level of detail regarding the diagnoses.

We found that in general, there were many attribute patterns, and the number of patterns was higher for women and for older individuals. In addition, the proportion of unique or rare patterns was very high. In the most detailed scenario 1, almost the entire dataset consisted of unique patterns (e.g. 95.4% unique patterns for individuals born in 1960). Although the frequency of patterns was higher for women and older individuals, the
proportion of unique patterns was slightly higher for men and younger individuals (women born in 1960 had 95.0% unique patterns, men born in 1960 had 96.1%, women born in 1980 had 96.9% and men born in 1980 had 98.7% unique patterns). In scenario 1, more than half of the statutory insured people had 20 or more unique attribute patterns (e.g. 71.0% for women born in 1960, 64.8% for 1970 and 64.5% for 1980).

4.2 Implications for re-identification potential and actual risk

Unique patterns in the dataset are the basis for the re-identification potential and associated re-identification risk. The basic individual risk measure that is linked to uniqueness of attribute patterns is 1/(frequency of that pattern in the population). This measures the risk without any further assumptions, e.g. about specific additional knowledge and goals of a data processor. The only assumption is that a data processor randomly knows at least one specific pattern and is able to find it in the dataset. For the DaTraV data, the probability of re-identifying one specific individual is then determined by its pattern frequency in the population (assuming further that our dataset roughly corresponds to the population as stated above). For scenario 1, the re-identification probability of any given individual in the dataset would therefore be very high.

Apart from this worst-case scenario, there are further assumptions to be tested and risk estimations to be compared. We have seen that for many records in the dataset, there is more than one attribute pattern. Thus, the assumption that a data processor knows one specific attribute pattern is quite strong. If he or she randomly knew a pattern that is very common in the dataset, re-identification would become more difficult.

The large number of unique patterns was mainly caused by the patterns of the diagnoses, since there are many different diagnoses possible. Following Elliot et al. 2002, we therefore have the case of many special uniques that are also unique on a subset of attributes. If a person has one rare diagnosis, the combinations with other diagnoses are rare or unique as well. Thus, such a person would have a substantial re-identification risk. However, as we have already noted above, the calculation of the SUDA score is difficult for the DaTraV data and has therefore not been applied for this paper.

4.3 Implications for statistical disclosure control

The preliminary results have emphasized the need for adequate statistical disclosure control. As a standard practice, the data processing centre in the DIMDI uses a minimum cell count rule (Deutsches Institut für Medizinische Dokumentation und Information 2016). The standard minimum cell count for each cell in tabular output should be 30. In well-founded exceptional cases, this minimum cell count can be decreased to no fewer than five. In the first step, risky cells with a cell count of less than the accepted minimum cell count are altered so that they satisfy the condition. In a following step, it is checked whether modified values can be recalculated exactly or approximately. If this is the case, further values are modified to ensure that the results of the recalculations also meet the minimum cell count rule. The standard minimum cell count of 30 makes it easier to handle values that can be approximately recalculated: The smaller the accepted
minimum number of cases, the more accurate are the results of approximate recalculations. The described procedure ensures that sensitive information is not communicated in detail (30 out of 70 million people are ~0.000043 %, or if one splits 70 million people into groups of 30, there will be a little more than 2.3 million groups).

4.4 Limitations and further steps

For this paper, we have only investigated outpatient diagnoses. However, quite a number of entitled institutions have access to hospital patient data. Therefore, inpatient diagnoses should be considered. In addition, data processors can have some knowledge about the medication of a patient, so this should be taken into consideration as well.

The risk scenarios we have investigated so far require that the data processor has detailed knowledge about each diagnosis. It is not unlikely that a data processor might just know the diagnoses very roughly. Thus, in future investigations, we plan to also consider coarser codings of the diagnoses. We expect that this will lead to a significant decrease of unique attribute patterns.

Since it can be assumed that data processors generally only have access to sections of the data or to similar data, we plan to conduct record linkage simulations with random samples in combination with result sets. This will allow us to investigate the effects of sampling, different selections of key variables, auxiliary information or various anonymisation methods.

Based on the achieved results we hope to find a way to better estimate the re-identification risk of the data by incorporating the re-identification potential, the intended utilisation and the settings in which the data and the result sets are distributed and handled.

References


