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

Hannah Lobert
Adam Giza
Dr. Jochen Dreß

Deutsches Institut für Medizinische Dokumentation und Information (DIMDI)

Work Session on Statistical Data Confidentiality
The Hague, October 2019
Outline

1. Background
2. Methods
3. Results
4. Discussion
Background

- Illness might predict or turn into a chronic illness
- Sensitive nature of health care data obliges institutions dealing with them to ensure confidentiality and prevent re-identification
- Assessment of re-identification risk necessary
  - re-identification potential
- Application to German statutory health insurance data
Background

- DaTraV data
- Data from ~70 million people with statutory health insurance in Germany
  - Data generated by physicians, hospitals, pharmacies
    - Invoice services
    - Compensate expenses between health insurance companies
    - Research or controlling → DIMDI
  - Socio-demographics, outpatient medication, inpatient and outpatient diagnoses, health insurance expenses
Background

- **DIMDI**
  - Trust centre
  - Data processing centre
    - Services to authorised institutions → data analysis, result sets
    - Confidentiality
    - Normally: aggregated result sets
    - Exception: analysis on-site
Modelling re-identification risk
Results of an expert survey

• Potentially suitable methods
  – Uniqueness measure (see e.g. Duncan et al. 2011)
  – Special uniques detection algorithm (Elliot et al. 2002)
  – Loglinear models (Skinner & Holmes 1998)
  – Record linkage (Domingo-Ferrer & Torra 2002)
  – Information theoretic approach (Antal et al. 2014)
  – Aggregated individual risk measures

• Importance of appropriate definition of risk scenarios (dependent on context)
Outline

1. Background
2. Methods
3. Results
4. Discussion
Application of the uniqueness measure

- **Basis** for many methods for modelling re-identification risk
- **Joint consideration** of variables/attributes
- Examination of *attribute patterns*
- **DaTraV data:** sample uniqueness → population uniqueness
- **Rare attribute values or rare attribute combinations** → higher re-identification risk
## Uniqueness measure

### Example

<table>
<thead>
<tr>
<th>No.</th>
<th>Sex</th>
<th>Born in</th>
<th>Diagnosis 1</th>
<th>Diagnosis 2</th>
<th>Diagnosis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>196X</td>
<td>E10 (diabetes), 1st half-year 20YY, confirmed diagnosis</td>
<td>J06 (upper respiratory infection), 2nd half-year 20YY, primary inpatient diagnosis</td>
<td>E10 (diabetes), 2nd half-year 20YY, secondary inpatient diagnosis</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>197X</td>
<td>K02 (caries), 1st half-year 20YY, confirmed diagnosis</td>
<td>A09 (gastro-intestinal disease), 1st half-year 20YY, confirmed diagnosis</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
</tbody>
</table>
## Uniqueness measure

### Example

<table>
<thead>
<tr>
<th>No.</th>
<th>Sex</th>
<th>Born in</th>
<th>Diagnosis 1</th>
<th>Diagnosis 2</th>
<th>Diagnosis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>196X</td>
<td>E10 (diabetes), 1st half-year 20YY, confirmed diagnosis</td>
<td>J06 (upper respiratory infection), 2nd half-year 20YY, primary inpatient diagnosis</td>
<td>E10 (diabetes), 2nd half-year 20YY, secondary inpatient diagnosis</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>197X</td>
<td>K02 (caries), 1st half-year 20YY, confirmed diagnosis</td>
<td>A09 (gastro-intestinal disease), 1st half-year 20YY, confirmed diagnosis</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
</tbody>
</table>
## Uniqueness measure

### Example

<table>
<thead>
<tr>
<th>No.</th>
<th>Sex</th>
<th>Born in</th>
<th>Diagnosis 1</th>
<th>Diagnosis 2</th>
<th>Diagnosis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>196X</td>
<td>E10 (diabetes), 1st half-year 20YY, confirmed diagnosis</td>
<td>J06 (upper respiratory infection), 2nd half-year 20YY, primary inpatient diagnosis</td>
<td>E10 (diabetes), 2nd half-year 20YY, secondary inpatient diagnosis</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>197X</td>
<td>K02 (caries), 1st half-year 20YY, confirmed diagnosis</td>
<td>A09 (gastro-intestinal disease), 1st half-year 20YY, confirmed diagnosis</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>198X</td>
<td>F32 (depression), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J45 (asthma), 2nd half-year 20YY, confirmed diagnosis</td>
<td>J20 (bronchitis), 1st half-year 20YY, confirmed diagnosis</td>
</tr>
</tbody>
</table>
**Uniqueness measure**

Selection of **key variables**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of birth</td>
<td>Numeric (4 digits)</td>
</tr>
<tr>
<td>Sex</td>
<td>1 – female, 2 – male</td>
</tr>
<tr>
<td>Documented ICD-code</td>
<td>Alphanumeric</td>
</tr>
<tr>
<td>Documentation period</td>
<td>Years, half-years, quarters</td>
</tr>
<tr>
<td>Sector</td>
<td>1 – outpatient, 2 – inpatient</td>
</tr>
</tbody>
</table>
Uniqueness measure

Definition of risk scenarios

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year of birth</td>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Documented ICD-codes</td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Documentation period</td>
<td>Quarters</td>
<td>Quarters</td>
<td>Half-years</td>
<td>Year</td>
</tr>
<tr>
<td>Qualification/type of diagnosis</td>
<td></td>
<td>All qualifications incorporated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sector</td>
<td></td>
<td></td>
<td>Separate consideration</td>
<td></td>
</tr>
<tr>
<td>Number of considered ICD-codes</td>
<td>4 diagnoses, with chronological order</td>
<td>2 diagnoses, with chronological order</td>
<td>2 diagnoses, with chronological order</td>
<td>2 diagnoses, without chronological order</td>
</tr>
</tbody>
</table>

→ Multiple patterns per person, analysis in batches
Outline

1. Background
2. Methods
3. Results
4. Discussion
## Uniqueness measure
Calculation of re-identification potential for selected cohorts

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Number of insurants</th>
<th>Number of attribute patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women born 1980</td>
<td>428,979</td>
<td>3,015,917,359</td>
</tr>
<tr>
<td>Men born 1980</td>
<td>351,513</td>
<td>540,947,536</td>
</tr>
<tr>
<td>Women born 1970</td>
<td>477,394</td>
<td>6,922,749,865</td>
</tr>
<tr>
<td>Women born 1960</td>
<td>544,015</td>
<td>19,050,562,519</td>
</tr>
<tr>
<td>Men born 1960</td>
<td>450,564</td>
<td>8,469,607,591</td>
</tr>
</tbody>
</table>

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

Fig. 1: Proportion of attribute patterns depending on the frequency of occurrence in the reporting year for persons with year of birth 1960.
Results: Uniqueness measure

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.
Results: Uniqueness measure

Fig. 3: Proportion of unique attribute patterns in the reporting year for each cohort.
Results: Uniqueness measure

Proportion of insured persons with at least X unique attribute patterns for each year of birth

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

Proportion of insured persons with at least X unique attribute patterns for each cohort

Men born in 1960
Men born in 1970
Men born in 1980

Women born in 1960
Women born in 1970
Women born in 1980

Fig. 5: Proportion of insured persons with at least X unique attribute patterns in the reporting year for women and men of birth cohorts 1960, 1970 and 1980.
Outline

1. Background
2. Methods
3. Results
4. Discussion
Summary
Results obtained so far

- Many attribute patterns
  - ... more for women than for men
  - ... slightly more for older cohorts
- Proportion of unique or rare patterns was very high
- Scenario 1:
  - Almost the entire dataset consisted of unique patterns
  - More than half of the statutory insured people had 20 or more unique attribute patterns
Implications for re-identification potential and actual risk

• Basic individual risk measure that is linked to uniqueness of attribute patterns is $1/(\text{frequency of that pattern in the population})$

• Assumption: Data processor randomly knows at least one specific pattern and is able to find it in the dataset → strong assumption with multiple patterns

• Unique patterns mainly caused by patterns of diagnoses (special uniques)
Implications for statistical disclosure control

- Minimum cell count rule
- Prevention of (approximate) recalculations/derivations
Limitations and further steps

- Inpatient and outpatient diagnoses
- Coarser definition of diagnoses
- Record linkage simulations

→ Based on the achieved results we hope to find a way to better estimate the re-identification risk with
  → the re-identification potential,
  → the intended utilisation
  → and the settings in which the data and the result sets are distributed and handled.
Thank you for your attention.

Hannah Lobert
Data analyst
Informationssystem Versorgungsdaten

Tel.  +49 (0)221 4724-375
Fax.  +49 (0)221 4724-340
E-Mail Hannah.Lobert@dimdi.de

DIMDI
Deutsches Institut für Medizinische Dokumentation und Information

Waisenhausgasse 36-38a
50676 Köln
www.dimdi.de
References


