Anonymizing Data to Promote Open Science

Prof. Dr. Fabian Prasser
Medical Informatics Lab
Berlin Institute of Health
Charité – Universitätsmedizin Berlin
Motivation

- Data sharing: Big data approaches in medical research, e.g.:
  - Precision medicine: high case numbers, detailed characterizations
  - Real-world evidence: secondary use, e.g. of routine clinical data for research
  - Collaborative research, e.g. data sharing across institutional boundaries

- Open science: Initiatives to improve the transparency, reproducibility and reusability of research results and research data, e.g.:
  - NIH Genomic Data Sharing Policy, Notice NOT-OD-14-124; 2014.
  - EMA Policy 0070 on Publication of Clinical Data for Medicinal Products for Human Use; 2014.

- Data protection requirements
Background: Anonymous data according to the GDPR

GDPR, Recital 26:

„The principles of data protection should apply to any information concerning an identified or identifiable natural person [...]“

„[...] To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, [...]to identify the natural person directly or indirectly [...]“

"[In doing so] all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments [...]"

+ Principles: such as data minimisation and storage limitation

Source: Regulation (EU) 2016/679 of the European parliament and the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
Background: Technical perspective

- Processing of personal (input) data in such a way that anonymous (output) data is produced. Example:

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZIP</th>
<th>Weight</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Male</td>
<td>81539</td>
<td>71</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>76</td>
<td>Male</td>
<td>81675</td>
<td>80</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>66</td>
<td>Male</td>
<td>81929</td>
<td>85</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>81</td>
<td>Male</td>
<td>80802</td>
<td>79</td>
<td>C25.1 Malignant neoplasm of body of pancreas</td>
</tr>
<tr>
<td>74</td>
<td>Male</td>
<td>81249</td>
<td>88</td>
<td>C25.2 Malignant neoplasm of tail of pancreas</td>
</tr>
<tr>
<td>71</td>
<td>Female</td>
<td>80335</td>
<td>69</td>
<td>C18.2 Malignant neoplasm of ascending colon</td>
</tr>
<tr>
<td>64</td>
<td>Female</td>
<td>80339</td>
<td>71</td>
<td>C18.4 Malignant neoplasm of transverse colon</td>
</tr>
<tr>
<td>69</td>
<td>Male</td>
<td>80637</td>
<td>75</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
<tr>
<td>55</td>
<td>Female</td>
<td>80638</td>
<td>77</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
<tr>
<td>61</td>
<td>Male</td>
<td>81667</td>
<td>67</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
</tbody>
</table>

Background: Trade-offs

No risk, no data

Legal side / protection

Scientific side / utility

Max. risk, all data

Bad science

Breaking the law

Data transformations

Risk and privacy models

General purpose and special purpose utility models

Source: Barth-Jones, Brussels Privacy Symposium, 2016
ARX: Features and applications

• **Comprehensive feature set:** „traditional“ approaches, Differential Privacy, game-theoretic methods, privacy-preserving machine learning.

• **Quite scalable:** Significantly outperforms related tools, used to anonymise datasets with billions of records.

• **Graphical tool:** Used in education and training by commercial and public institutions in several countries.

• **Wide range of applications:** Creation of open datasets and used to build anonymisation pipelines in several domains, e.g. by telecom providers, health insurances.

• **Industry friendly:** Integrated into several commercial products, core algorithms adopted by SAP HANA.

• **Open source:** More than 50,000 downloads.
ARX: Graphical frontend

[Image of a graphical interface with various windows and panels, likely related to ARX and its functionalities]

Source: https://arx.deidentifier.org/
Examples of guidelines and reports mentioning ARX

- UK Anonymization Network: The anonymization decision-making framework
- Dutch Ministry of Justice and Security: On statistical disclosure control technologies
- French Ministry of Social Affairs and Health: Health data Anonymity and risk of re-identification
- National Institute of Standards and Technology: Collaboration Space
- National Institute of Standards and Technology: De-Identifying Government Datasets
- European Medicine Agency: External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use
- European Union Agency for Network and Information Security: Privacy and data protection by design
- Finnish Ministry of Transport and Communications: Anonymization services: needs and implementation options
- Polish Ministry of Digitalization: Open data - Security standard
- Technology, Methods, and Infrastructure for Networked Medical Research: ToolPool health research
- Korean Ministry of Science and ICT: A research on de-identification technique for personal identifiable information
- Personal Data Protection Commission of Singapore: Guide to basic data anonymization techniques
- Office of the Australian Information Commissioner: The de-identification decision-making framework
Example: Anonymisation pipelines for the LEOSS registry

• LEOSS: A European registry capturing the clinical course of SARS-CoV-2 infected patients (https://leoss.net) established at University of Cologne
  • No informed consent necessary (anonymous reports).
  • Retrospective documentation after discharge / death.
  • All hospitalized patients including children eligible.
  • Immediate start after verification.

• Open Science approach
  • Registry hosted in a secure environment in Cologne.
  • Anonymous data is shared with researchers and the public.
  • Additional anonymisation procedures have been implemented for this purpose.

LEOSS: Development process

- Developed without access to primary data

LEOSS: Approach for the Public Use File

• Qualitative risk assessment
  • Comparison with risky variables mentioned in laws and guidelines

• Quantitative risk assessment following recommendations from the Opinion on Anonymisation Methods by the Article 29 Data Protection Working Party
  • Protection from Singling out, Linkability and Inference
  • Formal anonymization process deleting data based on mathematical models

• Withholding of records to ensure that protection holds also when data is updated repeatedly

• Modular extensions for Scientific Use File
  • Date shifting, categorization, suppression

## LEOSS: Schema of the Public Use File

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at diagnosis</td>
<td>Age of patient at time of diagnosis</td>
</tr>
<tr>
<td>Gender</td>
<td>Sex of patient</td>
</tr>
<tr>
<td>Month first diagnosis</td>
<td>Month of first confirmed diagnosis of COVID-19</td>
</tr>
<tr>
<td>Year first diagnosis</td>
<td>Year of first confirmed diagnosis of COVID-19</td>
</tr>
<tr>
<td>Uncomplicated phase</td>
<td>Indicates whether the patient has been through the uncomplicated phase of COVID-19</td>
</tr>
<tr>
<td>Complicated phase</td>
<td>Indicates whether the patient has been through the complicated phase of COVID-19</td>
</tr>
<tr>
<td>Critical phase</td>
<td>Indicates whether the patient has been through the critical phase of COVID-19</td>
</tr>
<tr>
<td>Recovery phase</td>
<td>Indicates whether the patient has been through the recovery phase of COVID-19</td>
</tr>
<tr>
<td>Vasopressors in complicated phase</td>
<td>Indicates whether vasopressors were used in the complicated phase</td>
</tr>
<tr>
<td>Vasopressors in critical phase</td>
<td>Indicates whether vasopressors were used in the critical phase</td>
</tr>
<tr>
<td>Invasive ventilation in critical phase</td>
<td>Indicates whether invasive ventilation was used in the critical phase</td>
</tr>
<tr>
<td>Superinfection in uncomplicated phase</td>
<td>Type of (if any) superinfection in uncomplicated phase</td>
</tr>
<tr>
<td>Superinfection in complicated phase</td>
<td>Type of (if any) superinfection in complicated phase</td>
</tr>
<tr>
<td>Superinfection in critical phase</td>
<td>Type of (if any) superinfection in critical phase</td>
</tr>
<tr>
<td>Symptoms in recovery phase</td>
<td>Symptoms (if any) in recovery phase</td>
</tr>
<tr>
<td>Last known patient status</td>
<td>Last known status</td>
</tr>
</tbody>
</table>

**LEOSS: Evaluation (1)**

- Pipeline based on the principle of “hiding in the crowd”
  - Anonymity is achieved by making sure that each record does not differ significantly from a larger group of records.
  - Counter-intuitive property: the greater the number of individuals included in the registry, the less information has to be removed to achieve the required degree of protection.

- Example: records released and case fatality rate

---

LEOSS: Evaluation (2)

- Example: descriptive statistics

Thank you for your attention!

Univ.-Prof. Dr. Fabian Prasser

Medical Informatics Lab
Berlin Institute of Health
Charité – Universitätsmedizin Berlin

https://www.bihealth.org/de/forschung/arbeitsgruppen/fabian-prasser/