Anonymizing Data to Promote Open Science

Prof. Dr. Fabian Prasser

Medical Informatics Lab
Berlin Institute of Health
Charité – Universitätsmedizin Berlin

Berlin Institute of Health
Charité & MDC

Aus Forschung wird Gesundheit

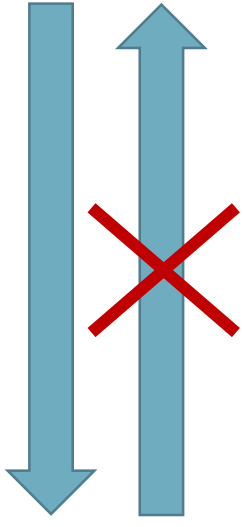
Berlin Institute of Health
Charité & MDC

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 Statement on Sharing Research Data, Notice NOT-OD-03-032; 2003.
  - 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

Personal data



Anonymous data

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:

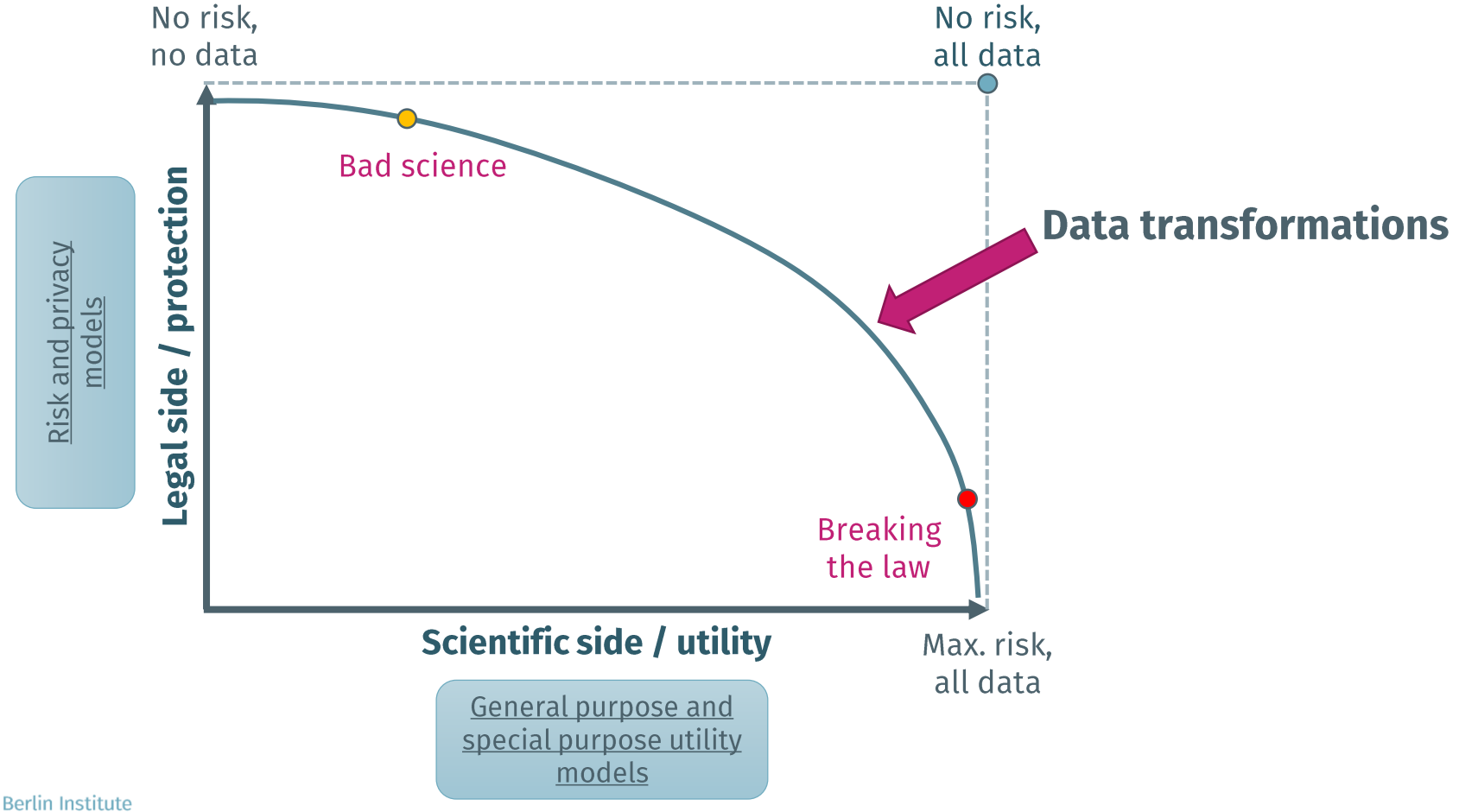
Age	Sex	ZIP	Weight	Diagnosis
55	Male	81539	71	C25.0 Malignant neoplasm of head of pancreas
76	Male	81675	80	C25.0 Malignant neoplasm of head of pancreas
66	Male	81929	85	C25.0 Malignant neoplasm of head of pancreas
81	Male	80802	79	C25.1 Malignant neoplasm of body of pancreas
74	Male	81249	88	C25.2 Malignant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

Age	Sex	ZIP	Weight	Diagnosis
72,0	Male	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	Male	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	Male	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
62,7	---	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	---	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	---	80***	[70, 80[	C18.- Malignant neoplasm of colon

- Sampling
- Aggregation
- Suppression
- Masking
- Categorization
- Generalisation

Source: Prasser et al. (2020) Flexible Data Anonymization Using ARX — Current Status and Challenges Ahead. J Software Pract Exper 50,7;1277-1304

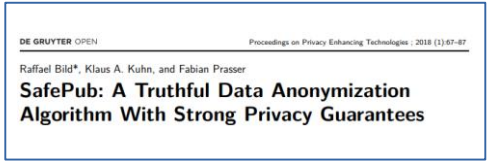
Background: Trade-offs

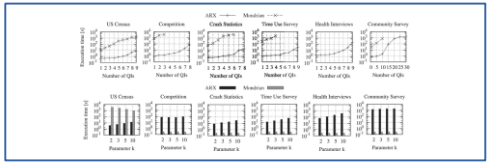


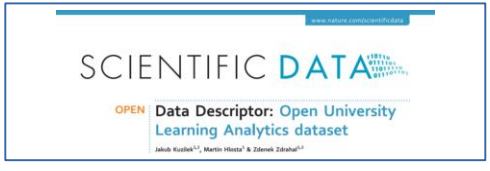
Source: Barth-Jones, Brussels Privacy Symposium, 2016

ARX: Features and applications

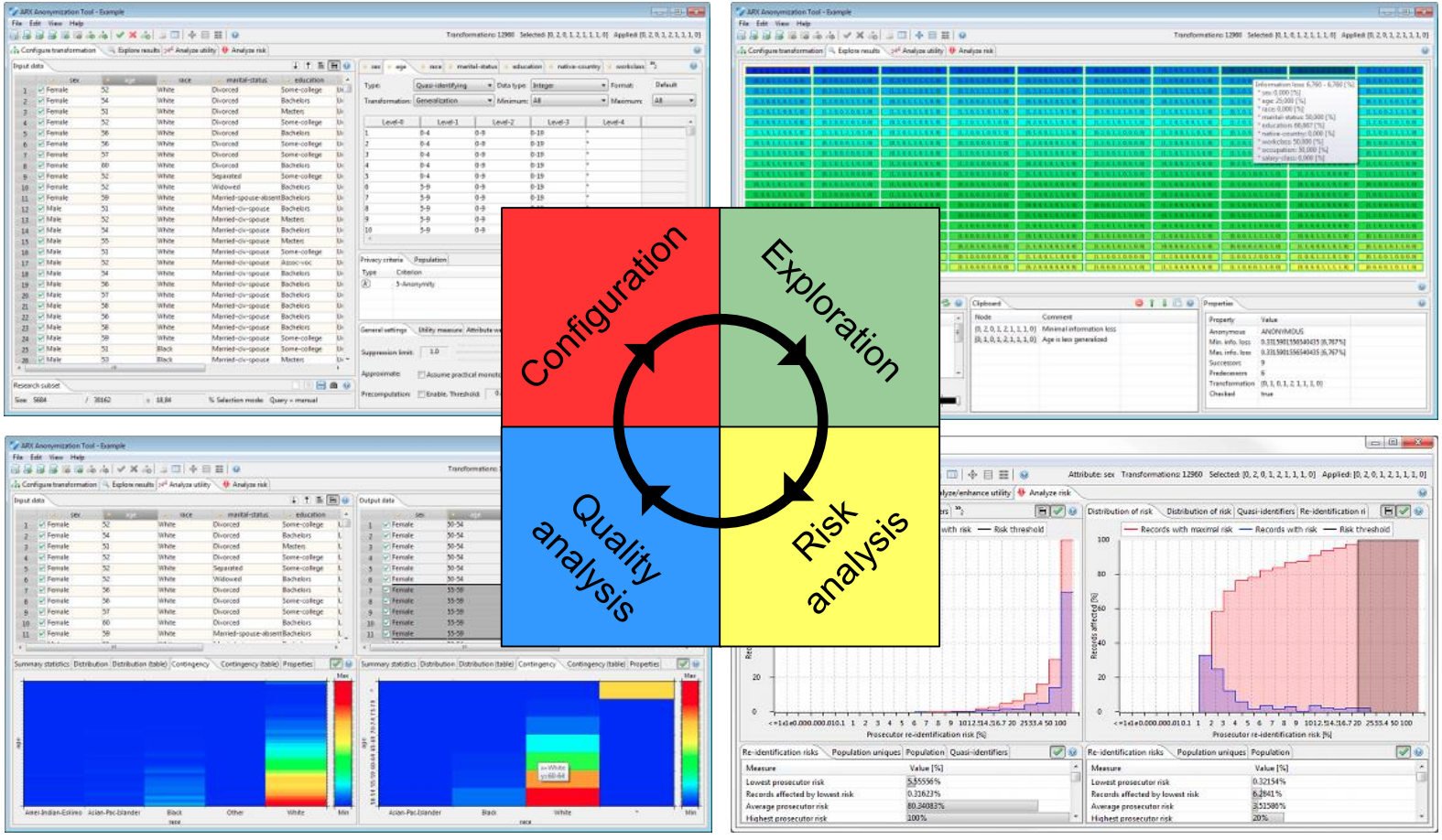
- **Comprehensive feature set**: „traditional“ approaches, Differential Privacy, game-theoretic methods, privacy-preserving machine learning.
- **Quite scalable**: Significantly outperforms related tools, used to anonymise datasets with billions of records.
- **Graphical tool**: Used in education and 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.





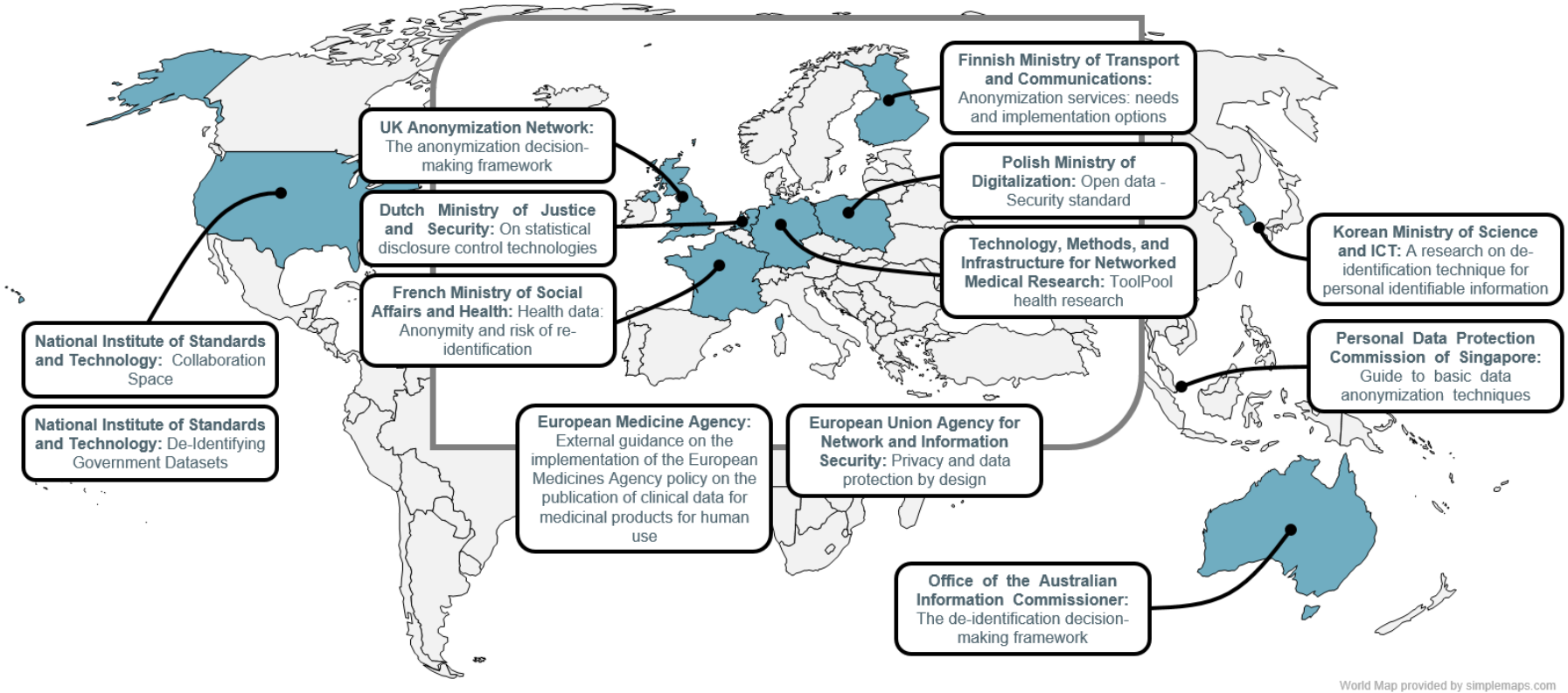


ARX: Graphical frontend



Source: https://arx.deidentifier.org/

Examples of guidelines and reports mentioning ARX



Example: Anonymisation pipelines for the LEOSS registry

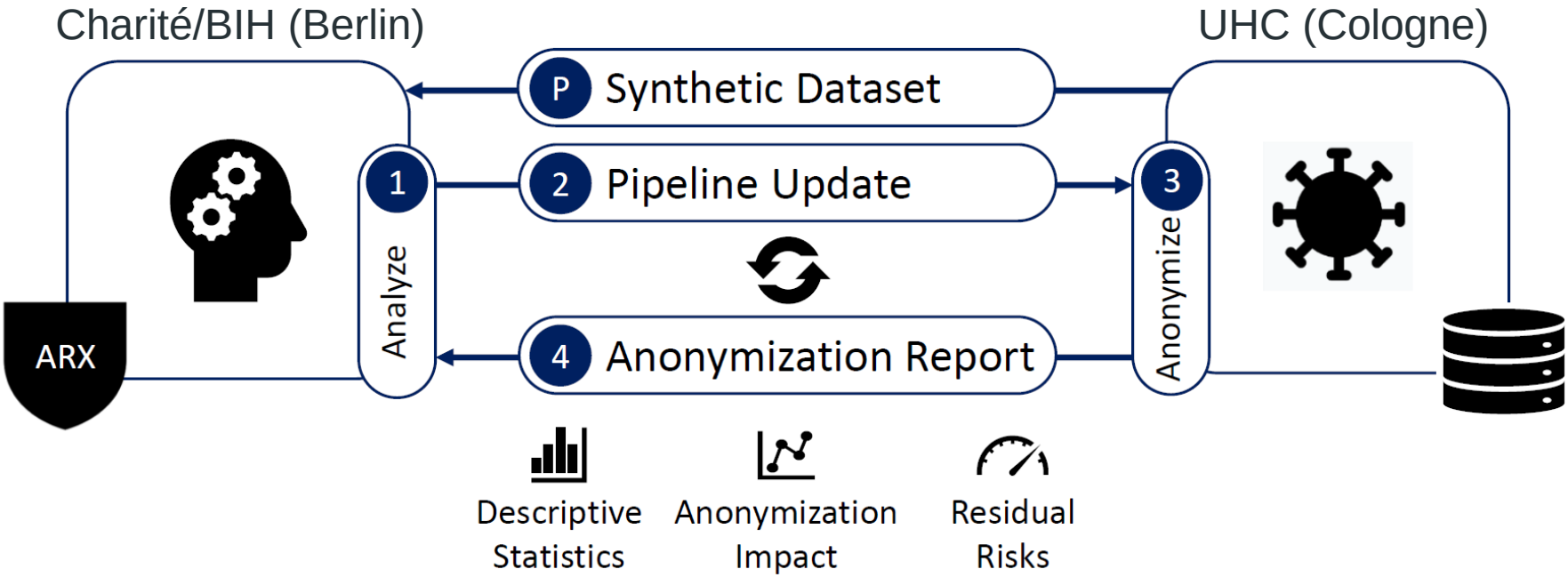
- LEOSS: A European registry capturing the clinical course of SARS-CoV-2 infected patients (https://leoss.net) established at University of Cologne
  - No informed consent necessary (anonymous reports).
  - Retrospective documentation after discharge / death.
  - All hospitalized patients including children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

Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

LEOSS: Development process

- Developed without access to primary data



Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

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

Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

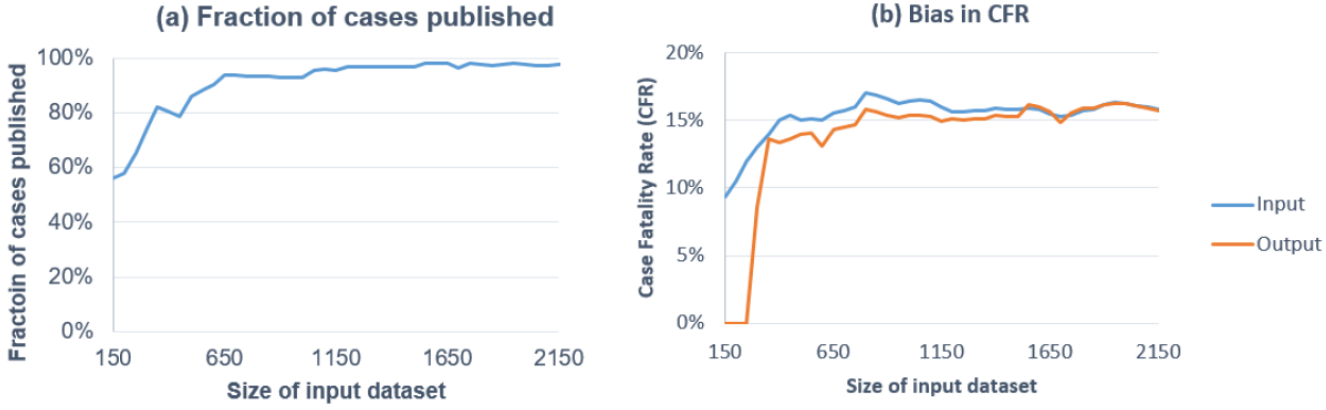
LEOSS: Schema of the Public Use File

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

Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

LEOSS: Evaluation (1)

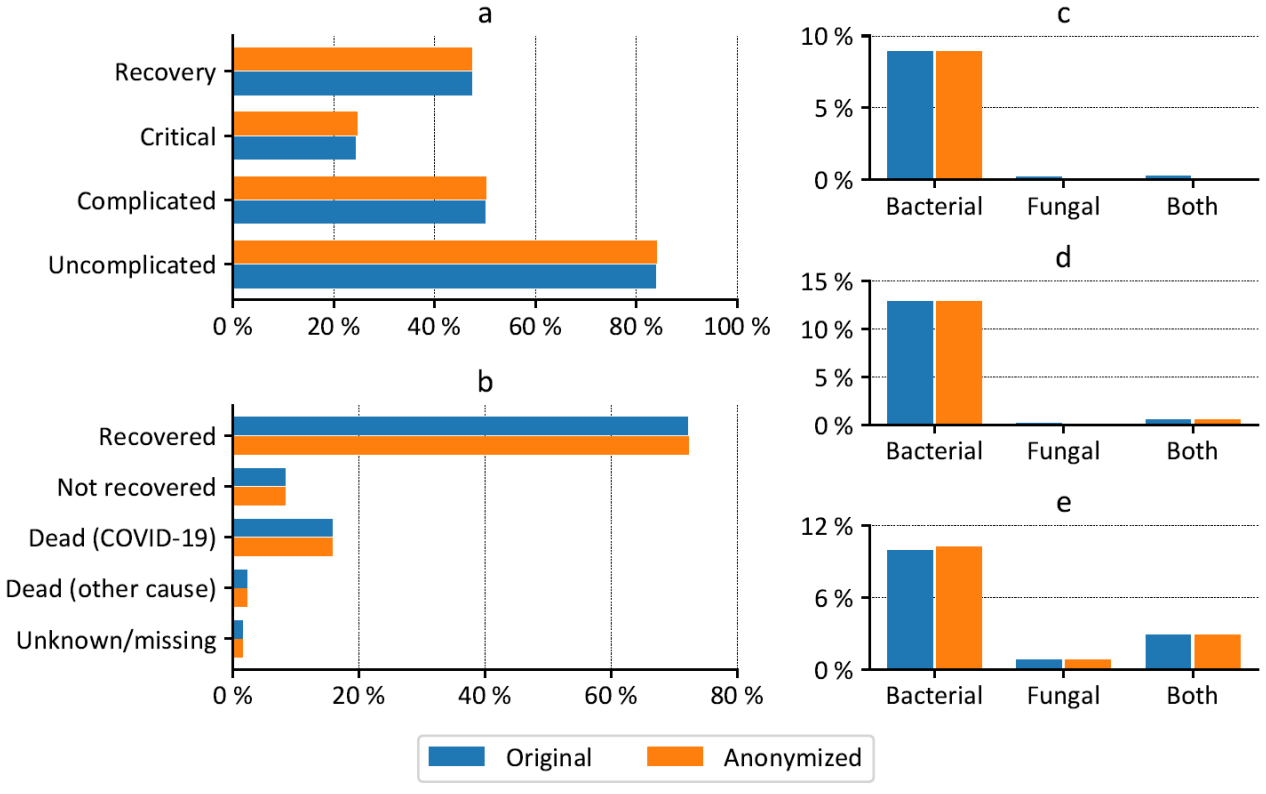
- Pipeline based on the principle of “hiding in the crowd”
  - Anonymity is achieved by making sure that each record does not differ significantly from a larger group of records.
  - Counter-intuitive property: the greater the number of individuals included in the registry, the less information has to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



Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

LEOSS: Evaluation (2)

- Example: descriptive statistics



Source: Jakob et al. (2020) Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific Data. In Press.

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/

Berlin Institute of Health
Charité & MDC

Aus Forschung wird Gesundheit