Lecture: Health Data Anonymization in Theory and Practice

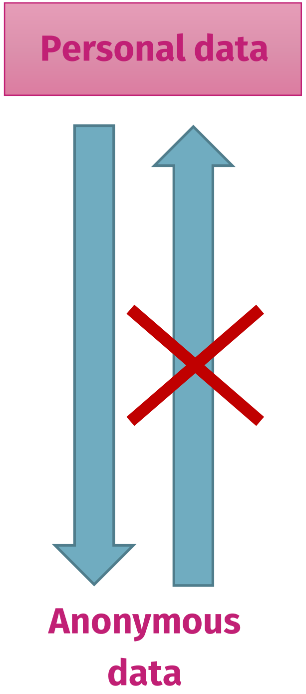
Prof. Dr. Fabian Prasser
Medical Informatics Group
Berlin Institute of Health @
Charité – Universitätsmedizin Berlin

BIH Berlin Institute of Health @Charité

Motivation: Data Sharing and Re-Use

- **Data-driven approaches in medical research**
  - 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
- **Initiatives to improve the transparency, reproducibility and reusability of research results and research data**
  - 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.
- **Increased citation rates**

EU General Data Protection Regulation (GDPR)



GDPR, Recital 26:

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

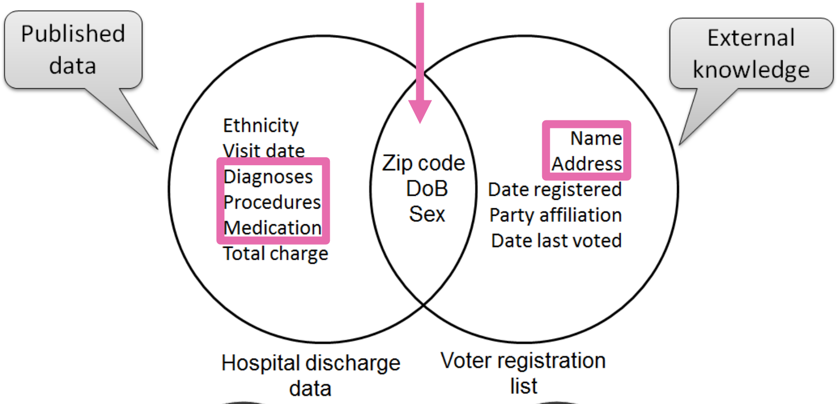
„[…] To determine whether a natural person is identifiable, **account should be taken of all the means reasonably likely to be used**, [..]to identify the natural person directly or indirectly […]“

"[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** […]"

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)

The Case of William Weld (1997): Linkage Attacks

Quasi-Identifiers



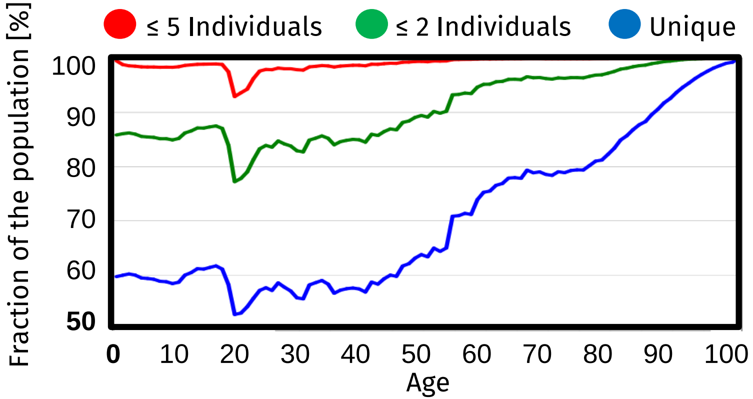
William Weld

Latanya Sweeney

Sources: Golle P. Revisiting the uniqueness of simple demographics in the US population. 5th ACM Workshop on Privacy in the Electronic Society, 2006, Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000, Image by By Gary Johnson from Taos, NM - BillWeld5x7 (2), CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=49683363

Uniqueness of Simple Demographics

87 % of the US population can be uniquely identified by the combination of DoB, Sex, ZIP Code



Sources: Golle P. Revisiting the uniqueness of simple demographics in the US population. 5th ACM Workshop on Privacy in the Electronic Society, 2006, Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000,

Re-Identification Revisited (2019)

Medical Data De-Identification Is Under Attack

David Talby Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program
Innovation

POST WRITTEN BY
David Talby
PhD, MBA, CTO at Pacific AI. Making AI, big data and data science solve real-world problems in healthcare, life science and related fields.

Forbes - Forbes Technology Council, 27.08.2019

The New York Times

Your Data Were 'Anonymized'? These Scientists Can Still Identify You

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.

The New York Times, 23.07.2019

"Anonymous" Data Won't Protect Your Identity

A new study demonstrates it is surprisingly easy to ID an individual within a supposedly incognito data set

Scientific American, 23.07.2019

nature COMMUNICATIONS

ARTICLE
https://doi.org/10.1038/s41467-019-10933-3 OPEN

Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher[1,2,3], Julien M. Hendrickx[1] & Yves-Alexandre de Montjoye[2,3]

Nature Communications, 23.07.2019

"[...] we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes."

BIH Berlin Institute of Health @Charité

Any Characteristic Can Make You Unique!

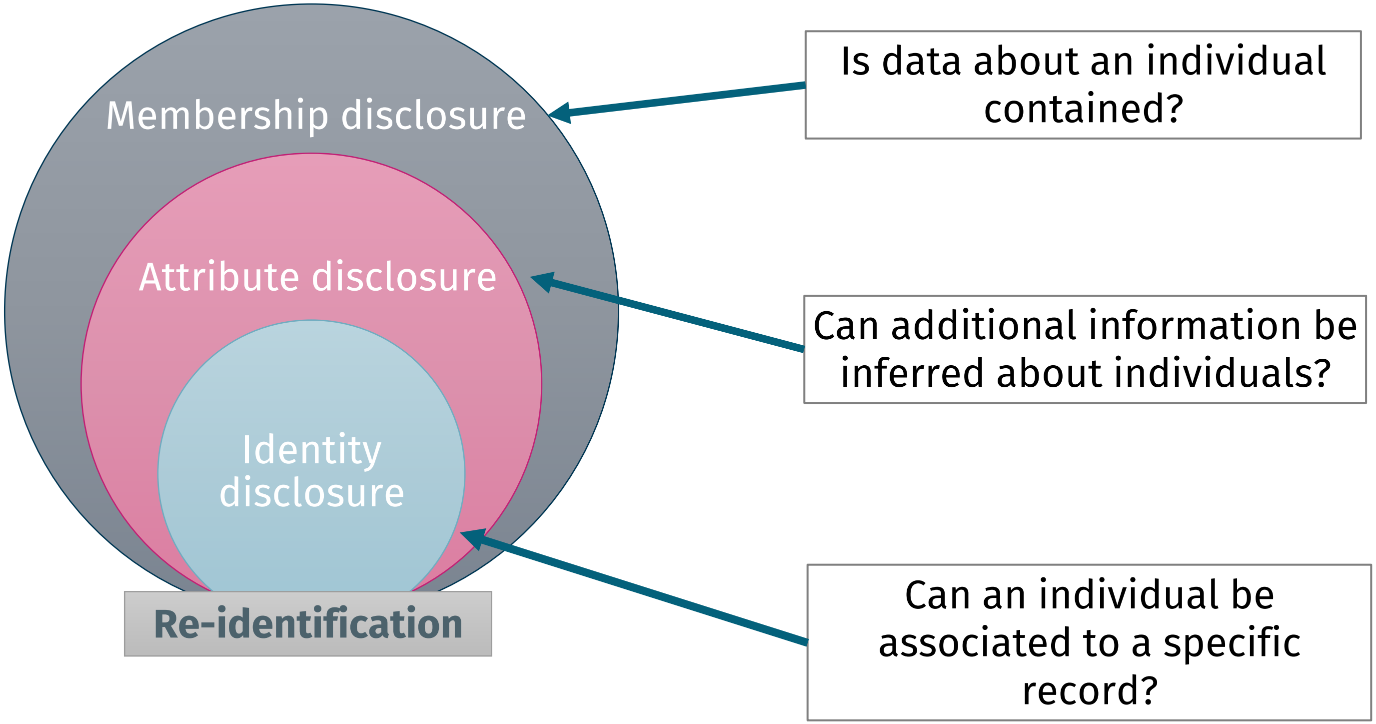
- **Demographic data** (Sweeney 1997; Golle 2006; El Emam 2008)
- **Diagnosis codes** (Loukides et al. 2010)
- **DNA (SNPs)** (Lin, Owen, & Altman 2004; Homer et al. 2008, Wang et al. 2009)
- **Pedigree structure** (Malin 2006)
- **Location visits** (Malin & Sweeney 2004, Golle & Partridge 2009)
- **Movie reviews** (Narayanan & Shmatikov 2008)
- **Search queries** (Barbaro & Zeller 2006)
- **Social network structure** (Backstrom et al. 2007, Narayanan & Shmatikov 2009)

But: Unique ≠ Identifiable!

Source: Bradley Malin. Challenges and Solutions for Data Privacy in Translational Research. 2011

BIH Berlin Institute of Health @Charité

(Re-)Identification: Types of Disclosure



(Re-)Identification: Example

Direct Id.		Quasi-Identifiers			Ins.	Sensitive Information
Name	Tel.	Age	Sex	ZIP	BMI	Diagnosis
*	*	[50,60)	M	92***	28.0	I47.1 Supraventricular tachycardia
*	*	[90,100)	M	94***	24.9	I50.1 Left ventricular failure
*	*	[80,90)	F	92***	26.7	I50.0 Congestive heart failure
*	*	[60,70)	F	94***	31.7	I47.1 Supraventricular tachycardia
*	*	[70,80)	F	92***	18.3	I47.0 Re-entry ventricular arrhythmia
*	*	[80,90)	F	92***	24.0	I50.0 Congestive heart failure
*	*	[80,90)	F	92***	28.1	I50.0 Congestive heart failure
*	*	[70,80)	M	94***	31.0	I47.0 Re-entry ventricular arrhythmia
*	*	[80,90)	M	94***	34.9	I50.0 Congestive heart failure
*	*	[60,70)	M	93***	32.3	I50.1 Left ventricular failure

1

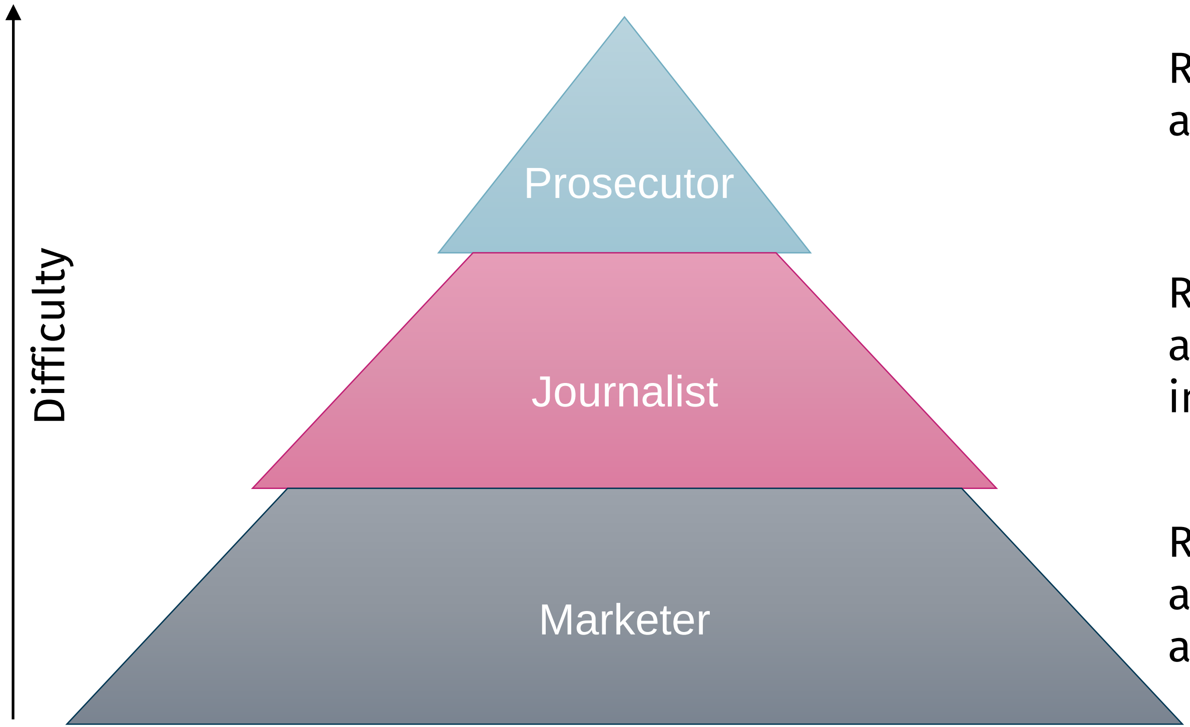
2

3

(1) Membership Disclosure
(2) Attribute Disclosure
(3) Identity Disclosure

Berlin Institute of Health @Charité

(Re-)Identification: Attacker Goals

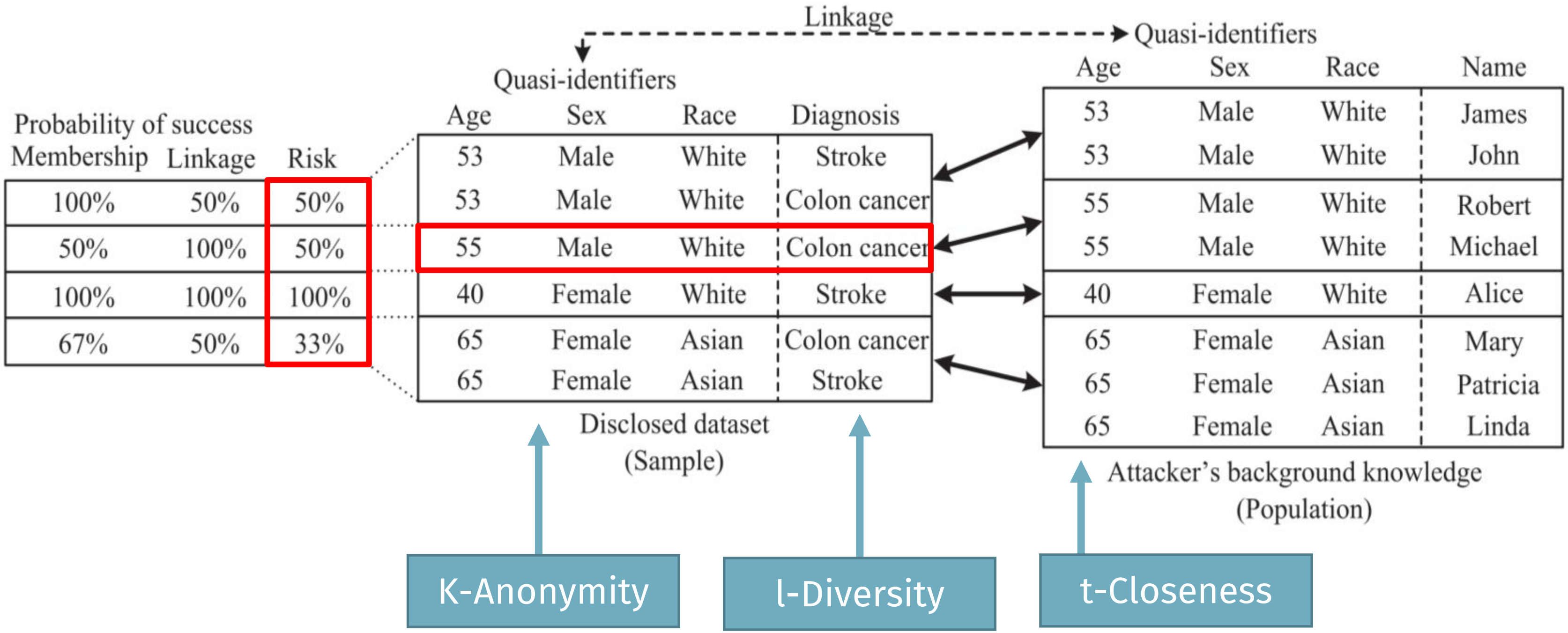


Re-identification of a specific individual

Re-identification of an unspecific individual

Re-identification of as many individuals as possible

Examples of Risk Models



Berlin Institute of Health @Charité

Anonymization: A Basic Example

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

Alter	Geschlecht	PLZ	Gewicht	Diagnose
55	Männlich	81539	71	C25.0 Bösartige Neubildung des Pankreas - Pankreaskopf
76	Männlich	81675	80	C25.0 Bösartige Neubildung des Pankreas - Pankreaskopf
66	Männlich	81929	85	C25.0 Bösartige Neubildung des Pankreas - Pankreaskopf
81	Männlich	80802	79	C25.1 Bösartige Neubildung des Pankreas - Pankreaskörper
74	Männlich	81249	88	C25.2 Bösartige Neubildung des Pankreas - Pankreasschwanz
71	Weiblich	80335	69	C18.2 - Bösartige Neubildung des Kolons - Colon ascendens
64	Weiblich	80339	71	C18.4 - Bösartige Neubildung des Kolons - Colon transversum
69	Männlich	80637	75	C18.7 - Bösartige Neubildung des Kolons - Colon sigmoideum
55	Weiblich	80638	77	C18.7 - Bösartige Neubildung des Kolons - Colon sigmoideum
61	Männlich	81667	67	C18.7 - Bösartige Neubildung des Kolons - Colon sigmoideum

Alter	Geschlecht	PLZ	Gewicht	Diagnose
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
62,7	---	80***	[70, 80[	C18.- Bösartige Neubildung des Kolons
62,7	---	80***	[70, 80[	C18.- Bösartige Neubildung des Kolons
62,7	---	80***	[70, 80[	C18.- Bösartige Neubildung des Kolons

Sampling

Aggregation

Suppression

Masking

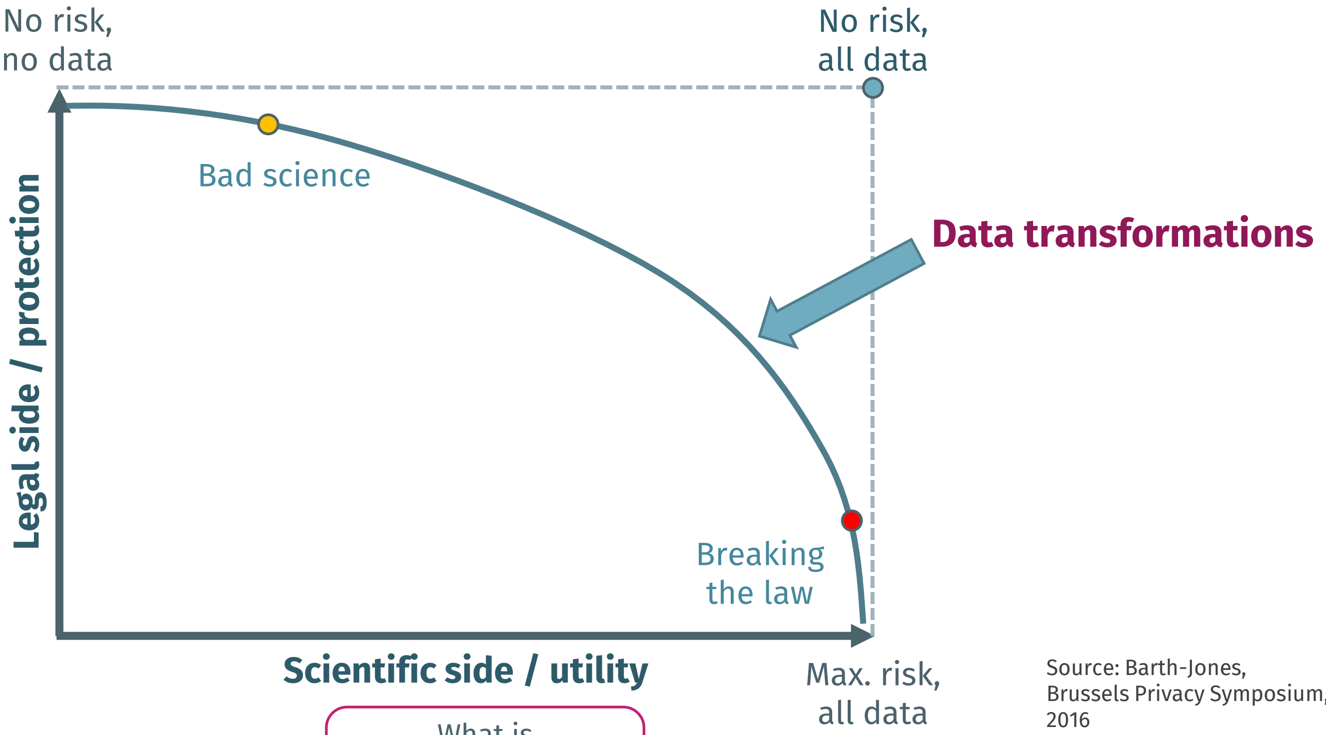
Categorization

Generalisation

k-Anonymity and (Ɛ, δ)-Differential Privacy

BIH Berlin Institute of Health @Charité

Trade-Offs in Anonymization



<u>What is</u>
- personal data?
- identification?
- acceptible risk?

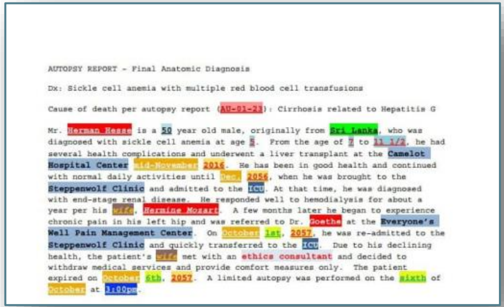
<u>What is</u>
- planned use?
- requirements?

Source: Barth-Jones, Brussels Privacy Symposium, 2016

Risk-Based Anonymization

Context: Purpose, recipient, types of data etc.

Source: https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface



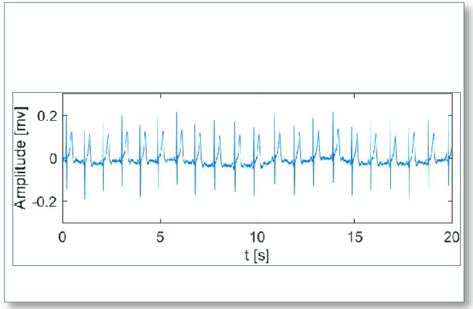
AUTOPSY REPORT – Final Anatomic Diagnosis

Dx: Sickle cell anemia with multiple red blood cell transfusions

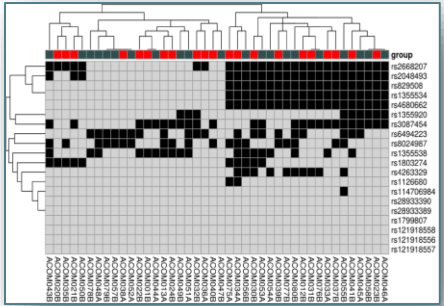
Cause of death per autopsy report (AU-01-23): Cirrhosis related to Hepatitis G

Mr. Herman Hesse is a 50 year old male, originally from Sri Lanka, who was diagnosed with sickle cell anemia at age 5. From the age of 7 to 11 1/2, he had several health complications and underwent a liver transplant at the Camelot Hospital Center mid-November 2016. He has been in good health and continued with normal daily activities until Dec. 2056, when he was brought to the Steppenwolf Clinic and admitted to the ICU. At that time, he was diagnosed with end-stage renal disease. He responded well to hemodialysis for about a year per his wife, Hermine Mozart. A few months later he began to experience chronic pain in his left hip and was referred to Dr. Goethe at the Everyone's Well Pain Management Center. On October 1st, 2057, he was re-admitted to the Steppenwolf Clinic and quickly transferred to the ICU. Due to his declining health, the patient's wife met with an ethics consultant and decided to withdraw medical services and provide comfort measures only. The patient expired on October 6th, 2057. A limited autopsy was performed on the sixth of October at 3:00pm.

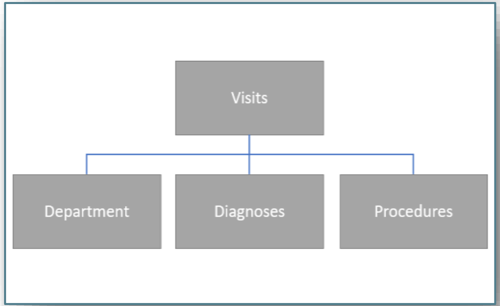
Source : https://scrubber.nlm.nih.gov/



Source : https://doi.org/10.1109/MeMeA.2018.8438751



Source : https://doi.org/10.2147/CCID.S176842



Source : https://www.g-drg.de/Datenlieferung_gem._21_KHEntgG

Onset of exposure	Yes	No	Total
20+ years***	339	53	392
0–19 years***	203	522	725
Total	542	575	1,117

Source : https://doi.org/10.1080/10937404.2012.678766

Berlin Institute of Health @Charité

Controlling the Context

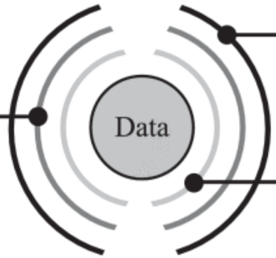
The Five Safes Framework

Safe People → Safe Projects → Safe Data → Safe Settings → Safe Outputs

Safe People, Safe Projects: Data Access Committees, IRBs, Data Use Agreements, AAI etc.

Safe Data, Safe Settings, Safe Outputs: Technical approaches

Controlled data sharing to keep oversight and prevent malicious attacks

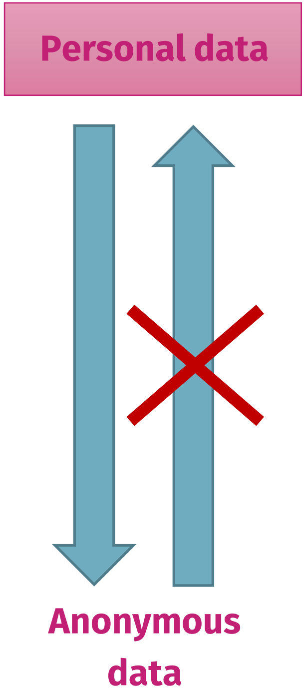


Organizational and legal measures to ensure trust

Tailored anonymization to present direct identification

Source: Desai, Ritchie, Welpton. (2016) Five Safes: designing data access for research.

Risk-Based Anonymization and 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 <u>reasonably likely</u> to be used**, [..]to identify the natural person directly or indirectly […]“

"[In doing so] all **<u>objective factors</u>, 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** […]"

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)

Overview of Available Tools

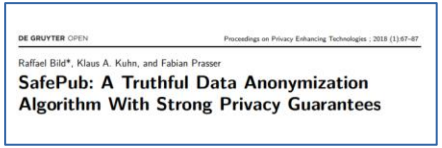
Tool	Institution	Country	Language(s)	Release	Last Update	License
µ-Argus	Centraal Bureau voor de Statistiek	Netherlands	C++, Java	1998	2021	EUPL
sdcMicro	Statistics Austria	Austria	R	2007	2021	GPL 2
Open Anonymizer	University of Vienna	Austria	Java	2008	2009	Unknown
CAT	Cornell University	USA	C++	2009	2014	Unknown
Tiamat	Purdue University	USA	Java	2009	Unknown	Unknown
UTD	The University of Dallas	USA	Java	2010	2012	GPL 2
Anon	University of Klagenfurth	Austria	Java	2012	Unknown	Unknown
ARX	BIH@Charité	Germany	Java	2012	2022	Apache 2
SECRETA	University of Peloponnes	Greece	C++, Qt	2013	Unknown	Unknown
Probabilistic Anonymization	University of Cyprus, Cyprus and Newcastle University, UK	Greece/UK	R	2018	2018	Unknown
µ-Ant	Center for Cybersecurity Research of Catalonia	Spain	Java	2019	2019	MIT
Amnesia	University of Thessaly	Greece	Java, JavaSript	2019	2022	BSD 3-Clause
PrioPrivacy	Research Studio Data Science	Austria	Java	2019	2021	Unknown

- Time focus around 2010 and 2020
- Most tools come from European institutions
- Most common programming languages are Java, C++, R
- Half of the tools identified are only publicly accessible to a limited extent
- Only a few tools are under permanent development
- Further research shows that only three of the tools (µ-Argus, ARX, sdcMicro) are used in real application scenarios.

Berlin Institute of Health @Charité

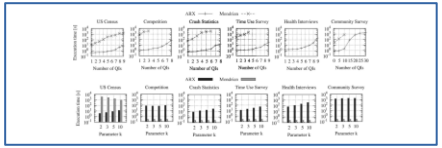
ARX: Features and Applications

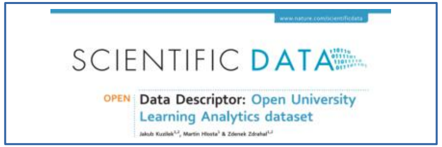
- **Comprehensive feature set:** „traditional“ approaches, Differential Privacy, game-theoretic methods, privacy-preserving machine learning.
- **Quite scalable:** Significantly outperforms related tools, used to anonymise datasets with billions of records.
- **Graphical tool:** Used in education and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


DE GRUYTER OPEN   Proceedings on Privacy Enhancing Technologies ; 2018 (1):67–87

Raffael Bild*, Klaus A. Kuhn, and Fabian Prasser

SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees




www.nature.com/scientificdata

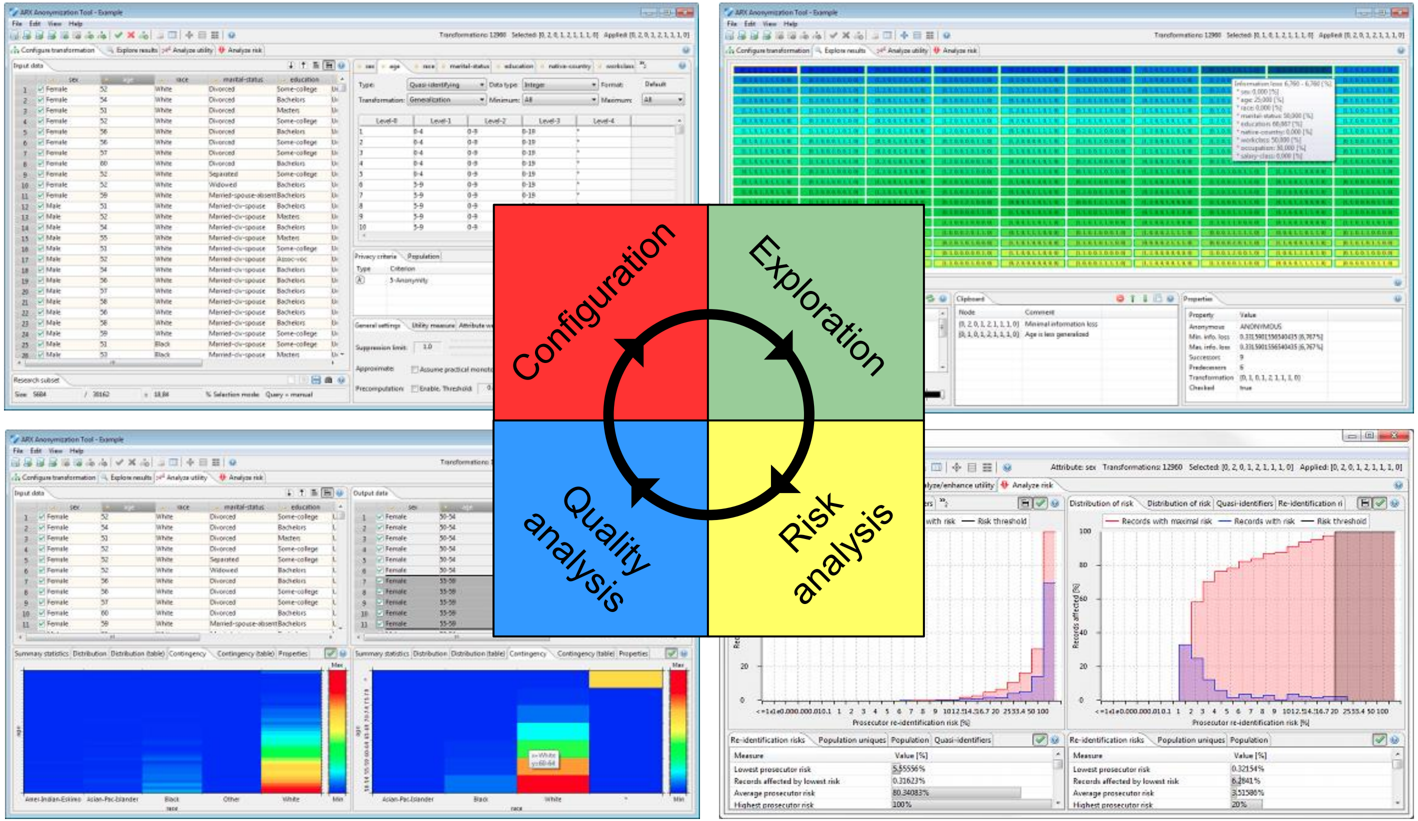
SCIENTIFIC DATA

OPEN Data Descriptor: Open University Learning Analytics dataset

Open for Innovation
KNIME

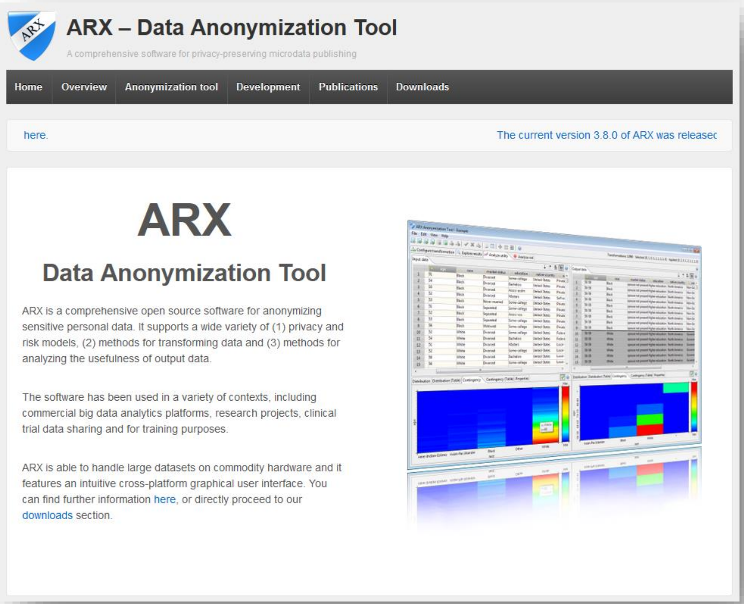
Berlin Institute of Health @Charité

ARX: Graphical Frontend

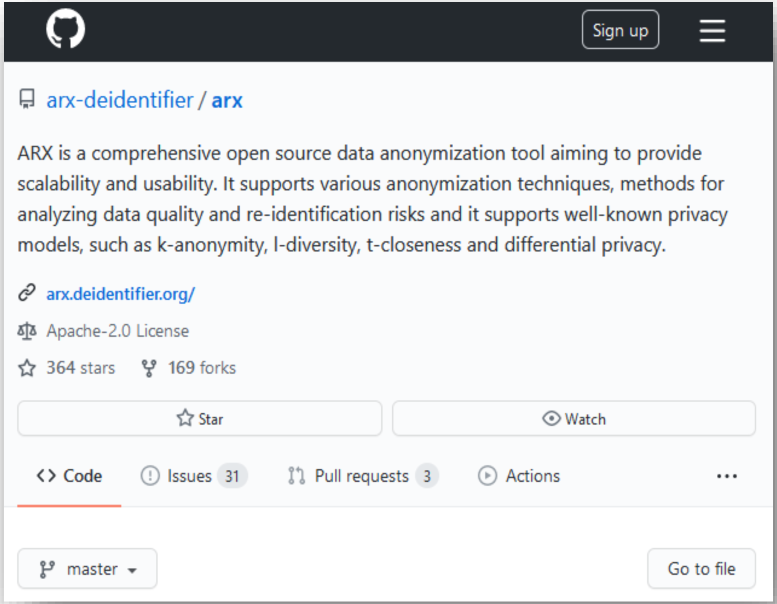


BIH Berlin Institute of Health @Charité

ARX: Open Source Project



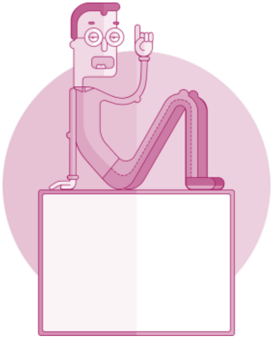
Comprehensive website



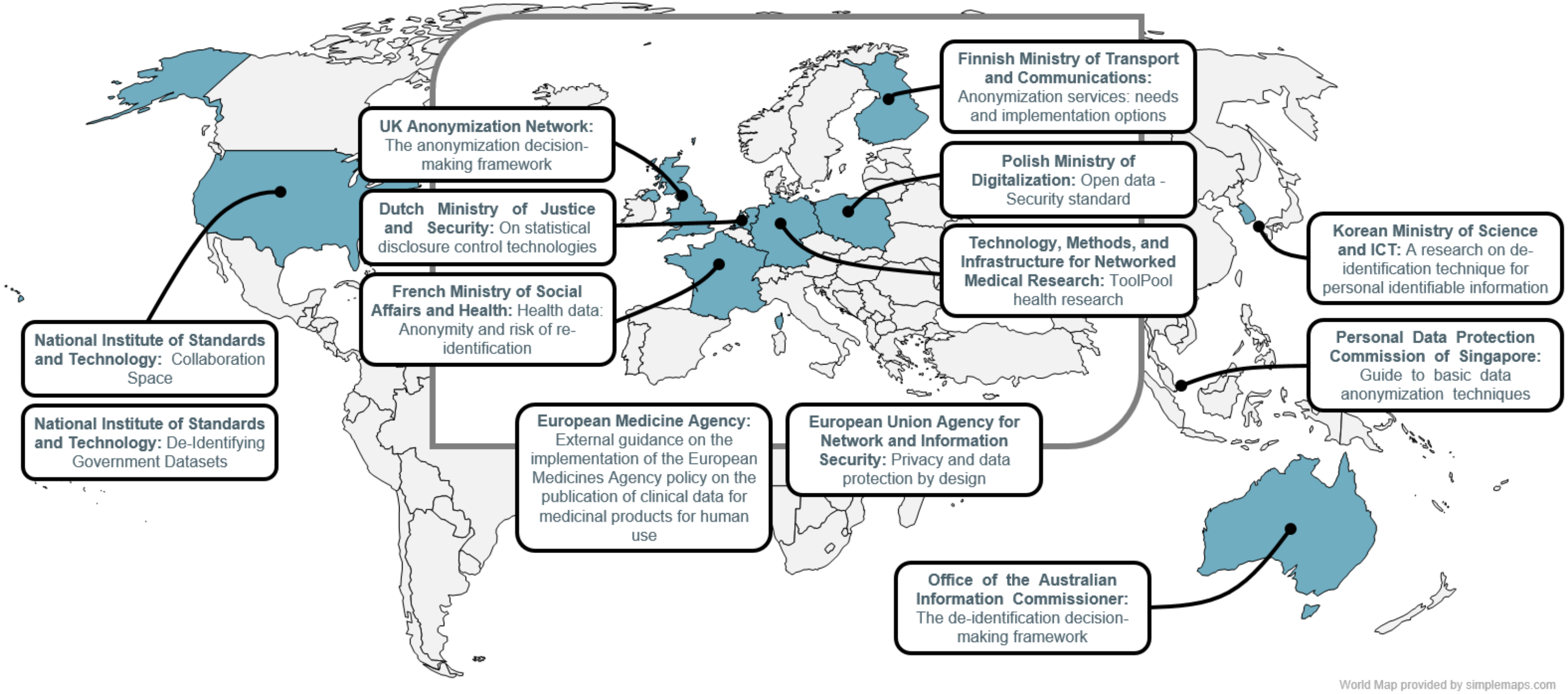
Community contributions

Examples of Guidelines Mentioning ARX (1)

- European Medicines Agency. EMA/90915/2016 – external guidance on the implementation of the European medicines agency policy on the publication of clinical data for medicinal products for human use; 2018.
- European Union Agency for Network and Information Security. Privacy and data protection by design; 2015.
- UKAN. The anonymisation decision-making framework; 2016.
- Office of the Australian Information Commissioner. The de-identification decision-making framework; 2017.
- French Ministry of Solidarity and Health. Health data: anonymity and risk of re-identification; 2015.
- Finnish Ministry of Transport and Communications. Anonymization services – requirements and implementation options; 2017.
- Personal Data Protection Commission of Singapore. Guide to basic data anonymisation techniques; 2018.
- Polish Ministry of Digitalization. Open data - Security standard; 2018.
- Dutch Ministry of Justice and Security. On statistical disclosure control technologies; 2018.
- Korean Ministry of Science and ICT. A research on de-identification technique for personal identifiable information; 2016.



Examples of Guidelines Mentioning ARX (2)



Berlin Institute of Health @Charité

Example: Anonymisation Pipelines for the LEOSS registry

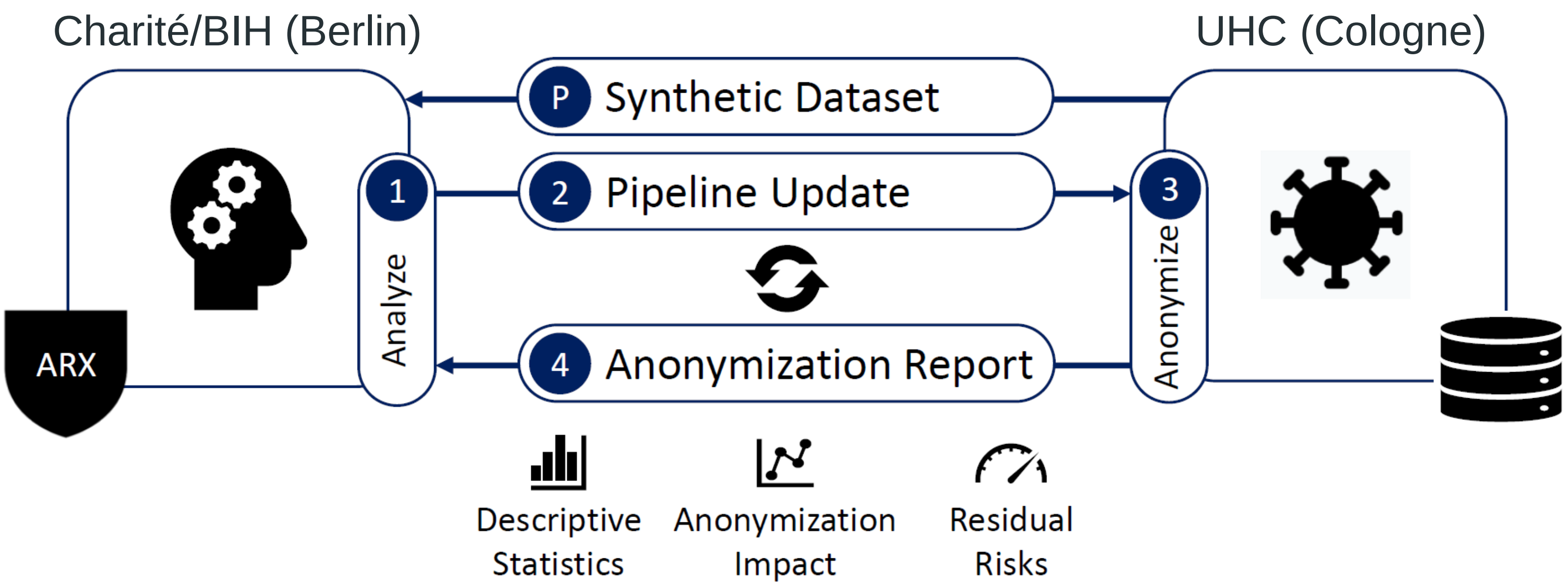
- **LEOSS: A European registry capturing the clinical course of SARS-CoV-2 infected patients (https://leoss.net) established at University of Cologne**
  - No informed consent necessary (anonymous reports).
  - Retrospective documentation after discharge / death.
  - All hospitalized patients including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

Berlin Institute of Health @Charité

LEOSS: Overview

- **Two types of datasets**
  - Public Use File with 16 variables available without restrictions.
  - Scientific Use Files with ≤605 variables available under data use contracts.
- **Two types of pipelines, built with ARX**
  - Two stages for the Public Use File
  - Ten stages for the Scientific Use File
- **Both pipelines were developed without access to primary data in close cooperation with the LEOSS Core Team in Cologne.**

LEOSS: Development Process



→ **Seven iterations over several weeks**

BIH Berlin Institute of Health @Charité

LEOSS: Result

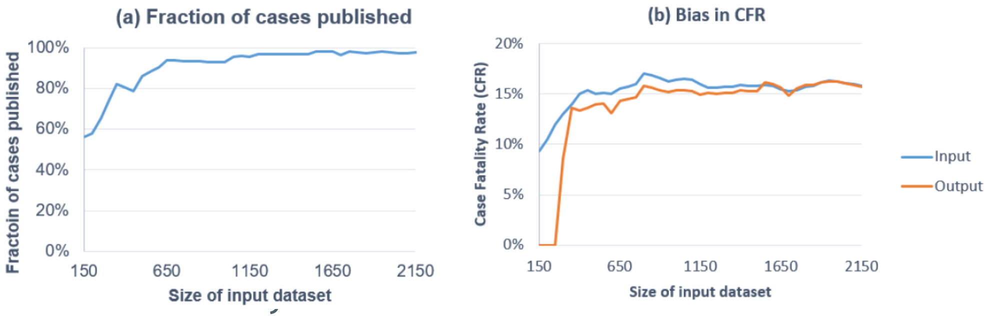
Variable	Description
Age at diagnosis	Age of patient at time of diagnosis
Gender	Sex of patient
Month first diagnosis	Month of first confirmed diagnosis of COVID-19
Year first diagnosis	Year of first confirmed diagnosis of COVID-19
Uncomplicated phase	Indicates whether the patient has been through the 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

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.

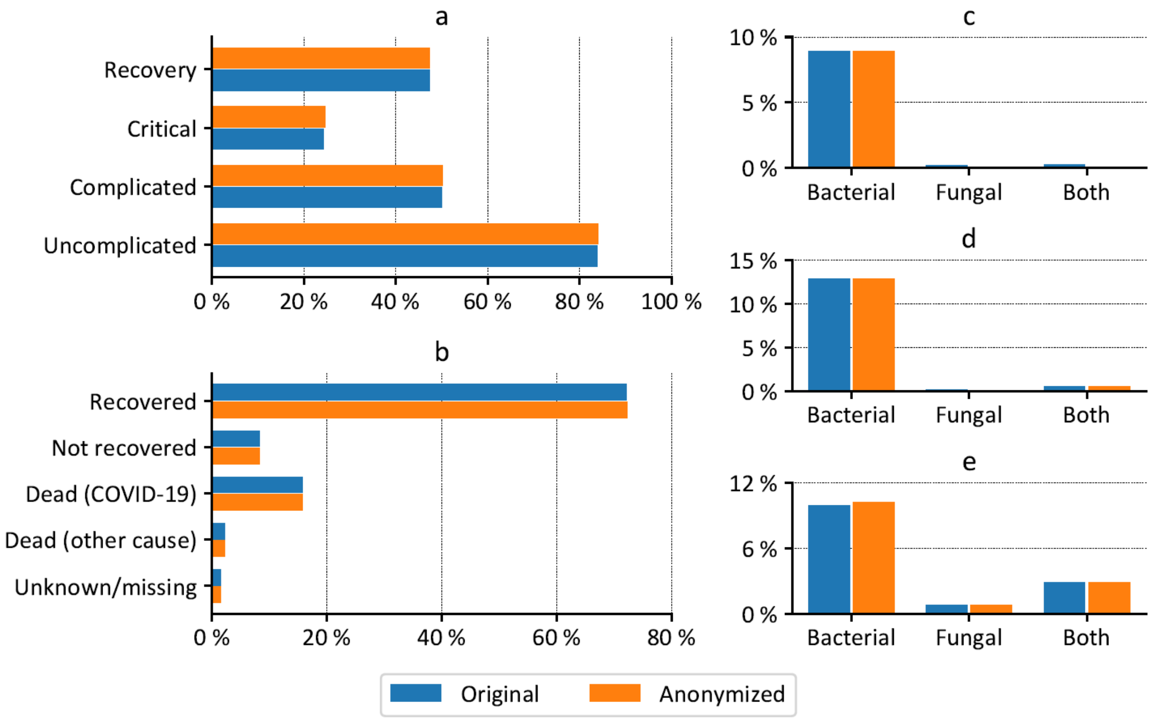
Example: records released and case fatality rate



→ Negligible imp

LEOSS: Evaluation (2)

Example: descriptive statistics



Berlin Institute of Health @Charité

LEOSS: Summary

Eight additional pipeline stages implement transformations for various modules of the Scientific Use File. Examples:

- Categorizing metric variables.
- Making timestamps relative.
- Grouping or suppressing sensitive variables.

  → Modules and stages can be activated dynamically to adjust to needs of different scientific / medical domains.

Overall approach

- Context-specific: adopted to the concrete dataset.
- Multiple layers of safeguards: qualitative + quantitative methods.
- Reliance on recommendations from laws and guidelines.
- Risk-based approach requires thorough documentation.

Questions?

Berlin Institute of Health @Charité