ARX: Datenanonymisierung in Theorie und Praxis

Prof. Dr. Fabian Prasser

Medical Informatics Group
Berlin Institute of Health @ Charité – Universitätsmedizin Berlin

March 2022

Berlin Institute of Health @Charité

Motivation: Data Sharing

Datengetriebene Ansätze in der medizinischen Forschung

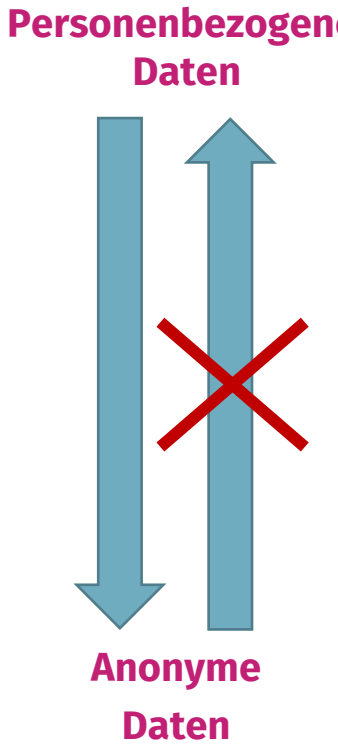
- Präzisionsmedizin: hohe Fallzahlen, detaillierte Charakterisierungen
- Real-World Evidence: Sekundärnutzung, z. B. von klinischen Routinedaten für die Forschung
- Kollaborative Forschung, z. B. gemeinsame Nutzung von Daten über institutionelle Grenzen hinweg

Initiativen zur Verbesserung der Transparenz, Reproduzierbarkeit und Wiederverwendbarkeit von Forschungsergebnissen und Forschungsdaten

- NIH-Erklärung zur gemeinsamen Nutzung von Forschungsdaten, Bekanntmachung NOT-OD-03-032; 2003.
- NIH-Richtlinie zur gemeinsamen Nutzung von Genomdaten, Bekanntmachung NOT-OD-14-124; 2014.
- EMA Policy 0070 on Publication of Clinical Data for Medicinal Products for Human Use; 2014.

Höhere Zitationsraten

Anonyme Daten: rechtliche Perspektive



DSGVO, Erwägungsgrund 26:

"Die Grundsätze des Datenschutzes sollten für **alle Informationen über eine bestimmte oder bestimmbare natürliche Person** gelten [...]"

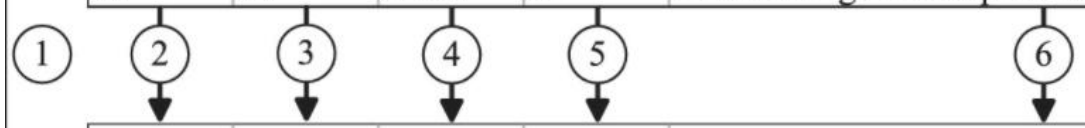
"[...] Um festzustellen, ob eine natürliche Person bestimmbar ist, sollten alle **Mittel berücksichtigt werden, die nach vernünftigem Ermessen eingesetzt werden können**, [...] um die natürliche Person direkt oder indirekt zu identifizieren [...]"

"[Dabei] sind alle **objektiven Faktoren, wie die Kosten und der Zeitaufwand für die Identifizierung**, unter Berücksichtigung der zum Zeitpunkt der Verarbeitung **verfügbaren Technologie und der technologischen Entwicklungen** zu berücksichtigen [...]"

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)

Anonymisierung: Beispiel

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



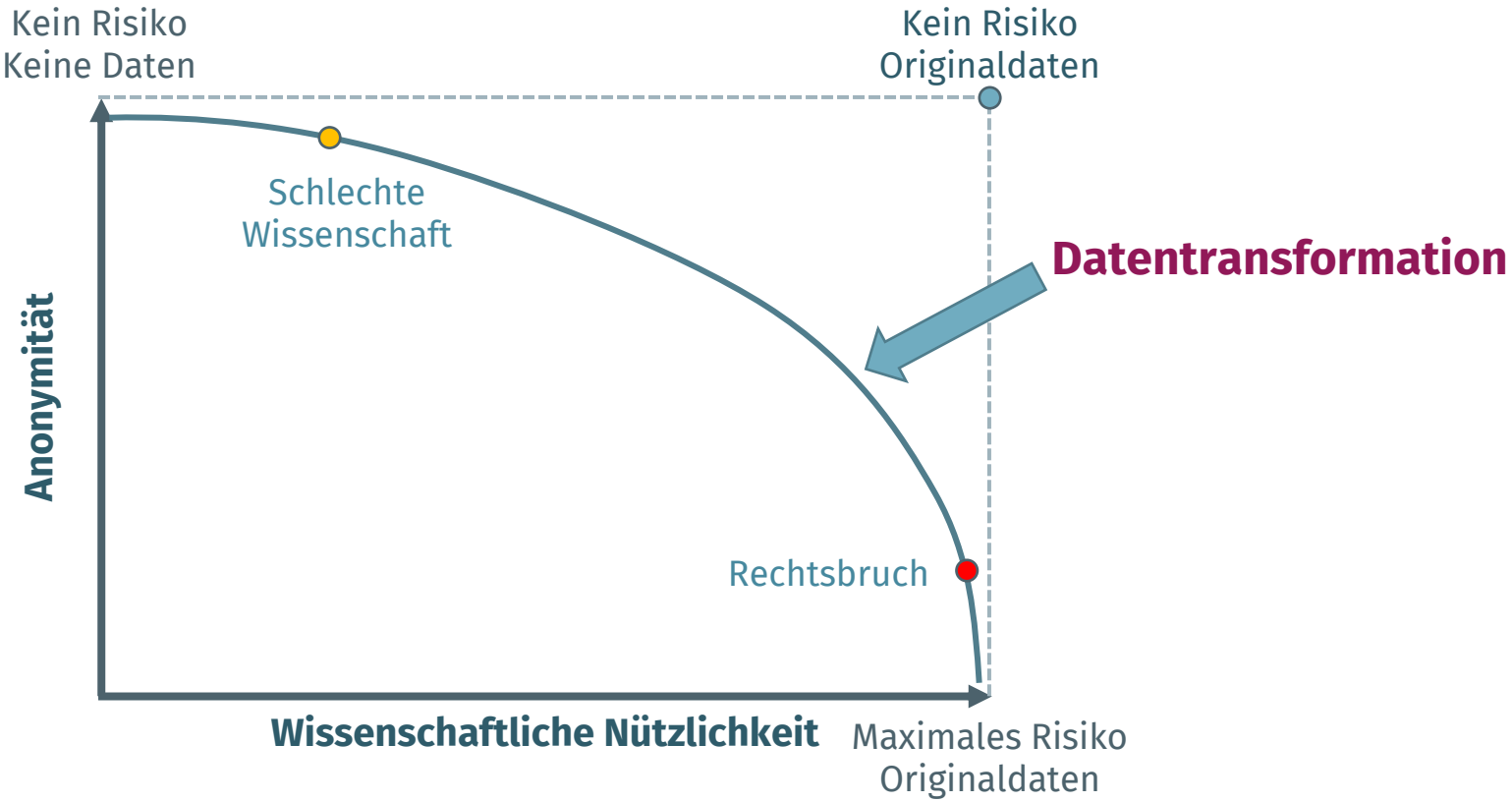
Age	Sex	ZIP	Weight	Diagnosis
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
72,0	M	81***	[80, 90[	C25.- Malignant neoplasm of pancreas
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm of colon
62,7	-	80***	[70, 80[	C18.- Malignant neoplasm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

1. Stichprobenziehung
2. Aggregation
3. Löschung
4. Maskierung
5. Kategorisierung
6. Generalisierung

Source: Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. Software: Practice and Experience. 2020 Jul;50(7):1277-304.

Anonyme Daten: technische Perspektive

Verschiedene Modelle und Grenzwerte

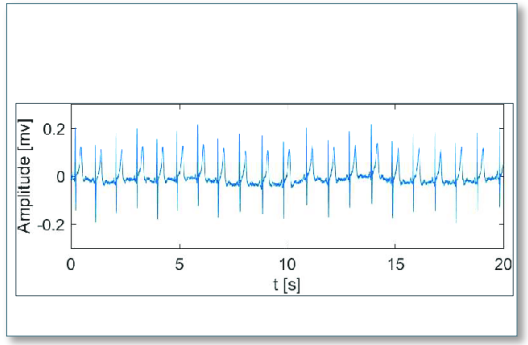


Source: Barth-Jones, Brussels Privacy Symposium, 2016

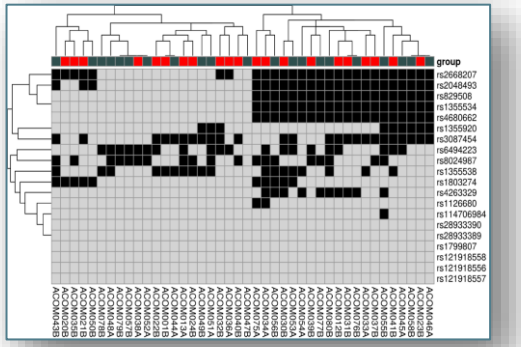
Risiken sind kontextabhängig!

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

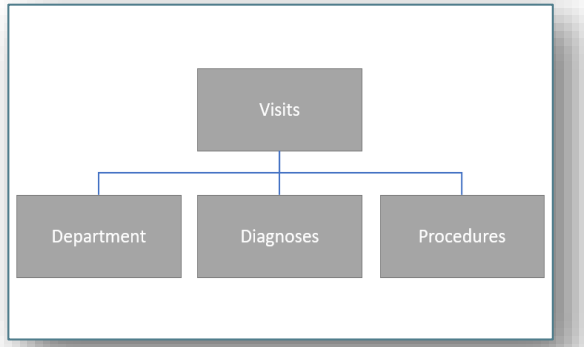
Source : https://scrubber.nlm.nih.gov/



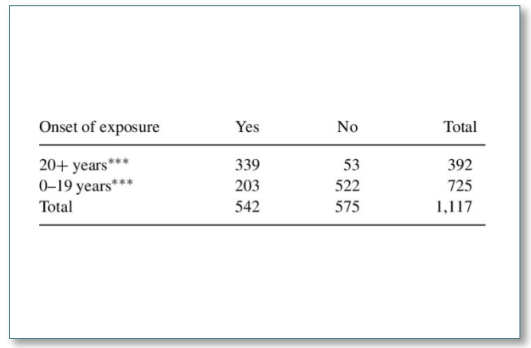
Source : https://doi.org/10.1109/MeMeA.2018.8438751



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

Marketer

Journalist

Prosecutor

Kontrolle des Kontexts

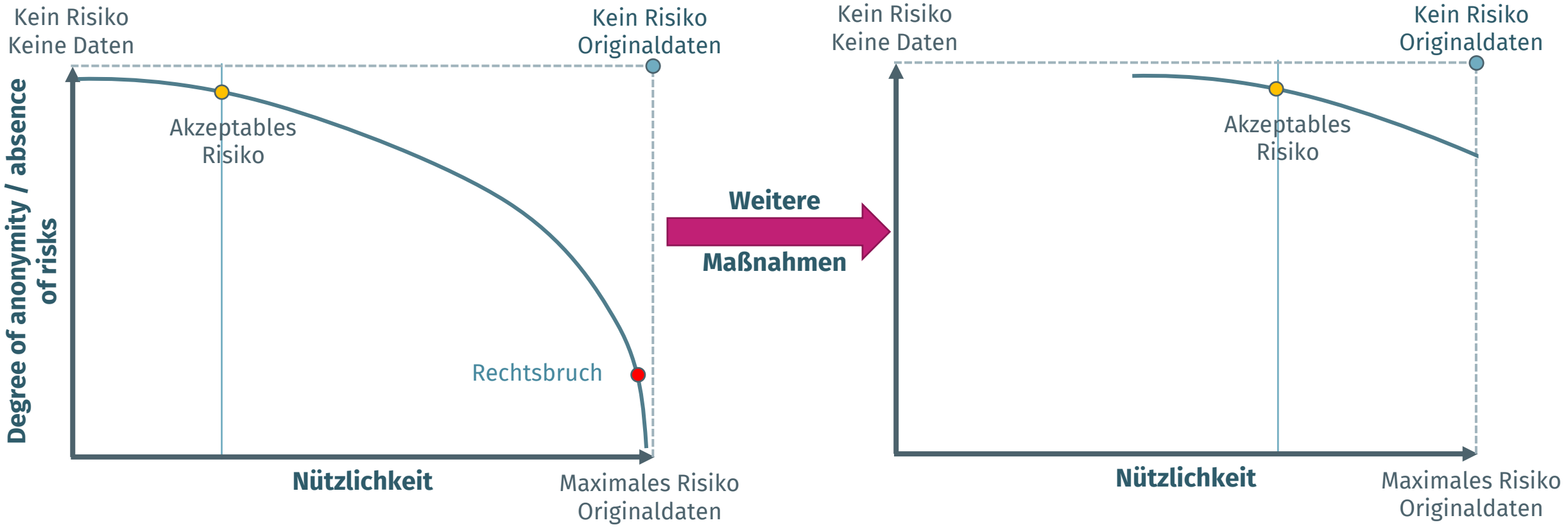
Das Five Safes Framework

Datenzugangskommittees, Ethikkommissionen, Datennutzungsvereinbarungen, AAI etc.

Technische Ansätze

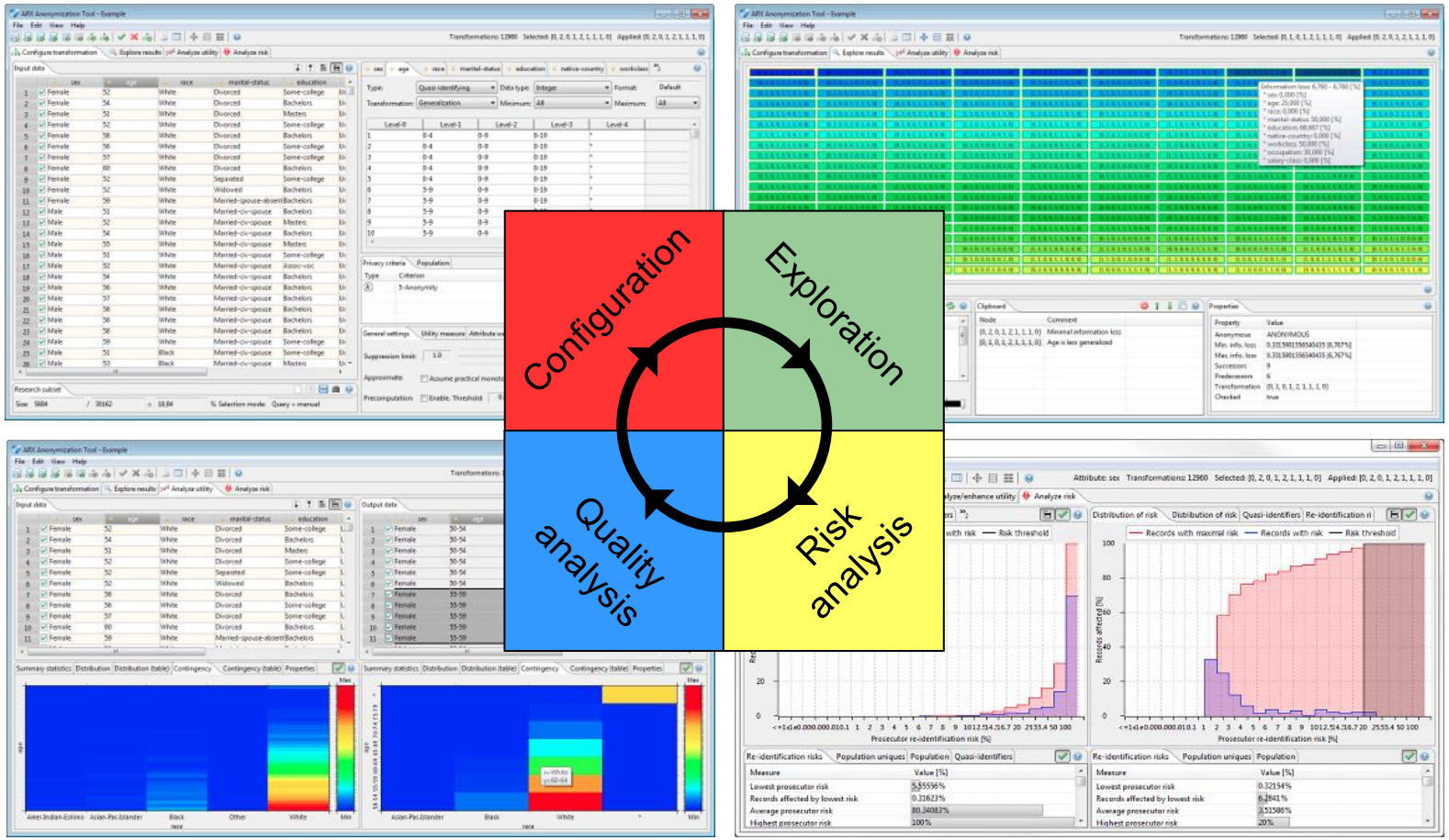
Source: Desai, Ritchie, Welpton. (2016) Five Safes: designing data access for research.

Bedeutung weiterer Maßnahmen



Inspired by: Barth-Jones, Brussels Privacy Symposium, 2016

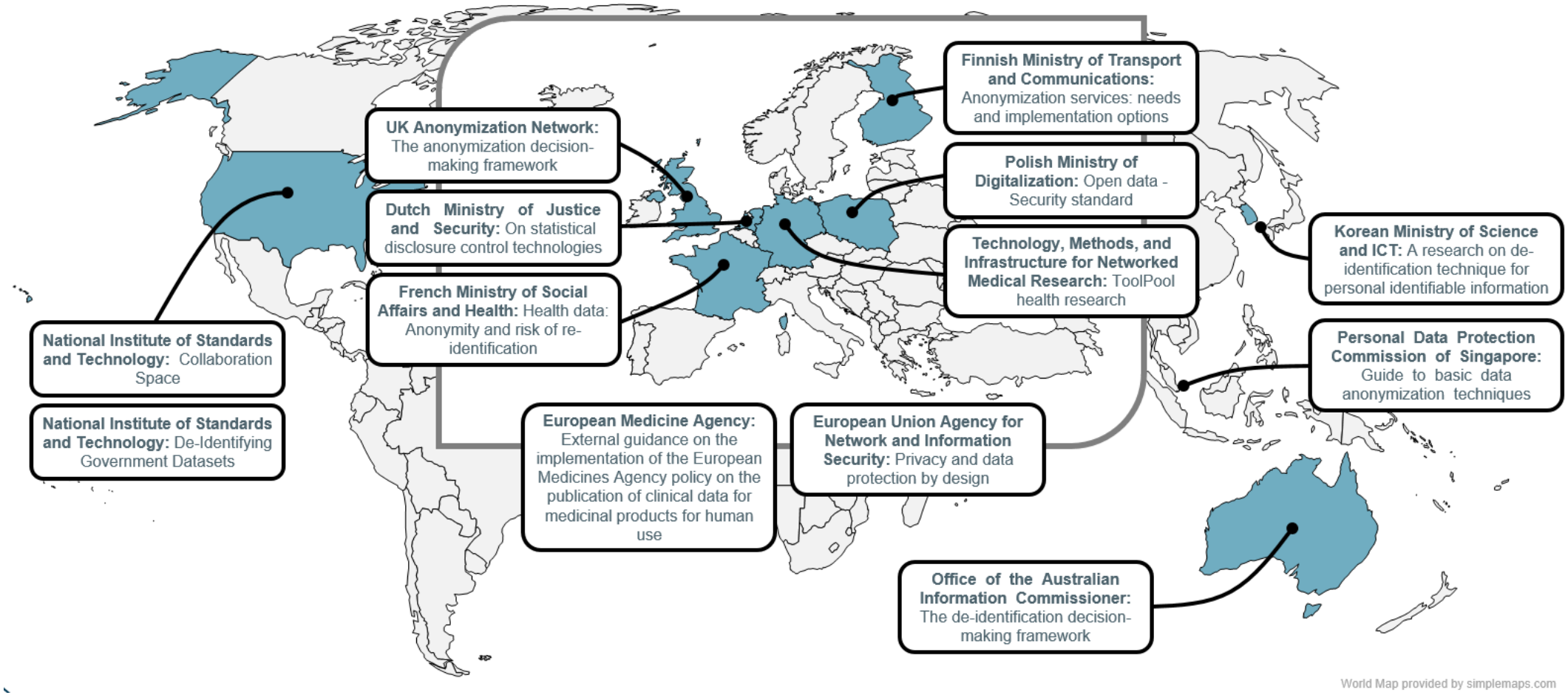
ARX Data Anonymization Tool



https://arx.deidentifier.org

Source: Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. Software: Practice and Experience. 2020 Jul;50(7):1277-304.

ARX: Impact



Source: Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—Current status and challenges ahead. Software: Practice and Experience. 2020 Jul;50(7):1277-304.

Anwendungsbeispiel: LEOSS (1)

LEOSS: Ein europäisches Register zur Erfassung des klinischen Verlaufs von mit SARS-CoV-2 infizierten Patienten, das an der Universität zu Köln eingerichtet wurde

- Keine informierte Zustimmung erforderlich (anonyme Berichte)
- Retrospektive Dokumentation nach Entlassung / Tod
- Alle hospitalisierten Patienten, auch Kinder, können teilnehmen
- Sofortiger Start nach Verifizierung

Open Science-Ansatz

- Das Register wird in einer sicheren Umgebung in Köln gehostet
- Anonyme Daten werden an Forscher und die Öffentlichkeit weitergegeben
- Zusätzliche Anonymisierungsverfahren werden von Fall zu Fall angewandt

Source: Pilgram L, Schons M, Jakob CE et al. The COVID-19 Pandemic as an Opportunity and Challenge for Registries in Health Services Research: Lessons Learned from the Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS). Das Gesundheitswesen. 2021 Nov;83(S 01):S45-53.

Anwendungsbeispiel: LEOSS (2)

Risikoanalyse

- Vergleich der Daten mit "risikobehafteten" Variablen, die in Gesetzen und Leitlinien genannt werden.
- Kontextspezifische Bewertung des mit den einzelnen Variablen verbundenen Identifizierungsrisikos unter Berücksichtigung von Replizierbarkeit, Verfügbarkeit und Unterscheidbarkeit ihrer Werte

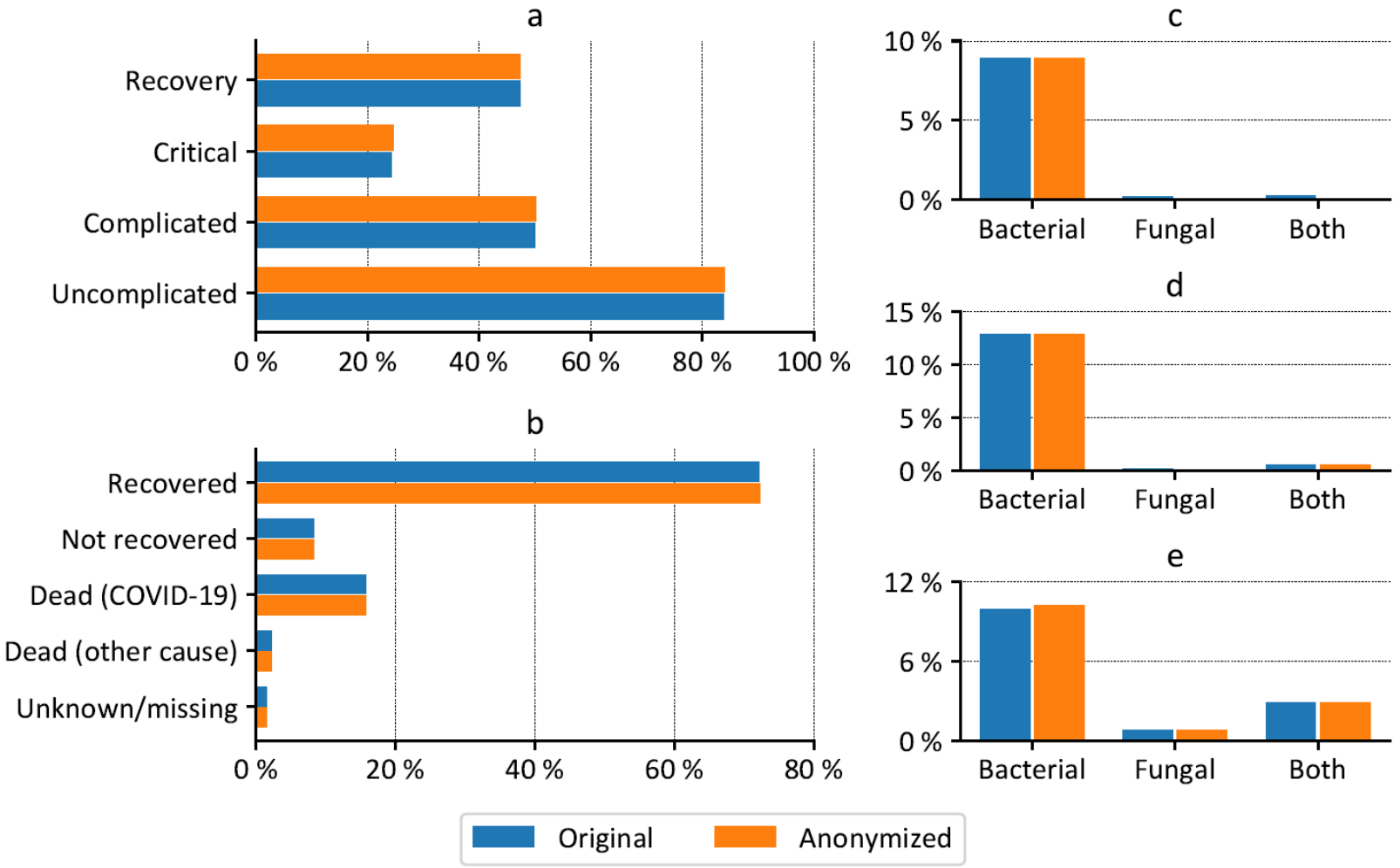
Formale Anonymisierung mit ARX

Auf der Grundlage der Stellungnahme zu Anonymisierungsmethoden der Artikel-29-Datenschutzgruppe (heute: Europäischer Datenschutzausschuss):

- Aussonderung: die Möglichkeit, einige oder alle Datensätze zu isolieren, die eine Person im Datensatz identifizieren.
- Verknüpfbarkeit: die Möglichkeit, mindestens zwei Datensätze zu verknüpfen, die dieselbe betroffene Person oder eine Gruppe von betroffenen Personen betreffen.
- Inferenz: die Möglichkeit, mit erheblicher Wahrscheinlichkeit den Wert eines Attributs aus den Werten einer Reihe anderer Attribute abzuleiten.

Source: Pilgram L, Schons M, Jakob CE et al. The COVID-19 Pandemic as an Opportunity and Challenge for Registries in Health Services Research: Lessons Learned from the Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS). Das Gesundheitswesen. 2021 Nov;83(S 01):S45-53.

Anwendungsbeispiel: LEOSS (3)



Source: Jakob CE, Kohlmayer F, Meurers T, Vehreschild JJ, Prasser F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Scientific data. 2020 Dec 10;7(1):1-0.

Danke für Ihre Aufmerksamkeit!

mi.bihealth.org

BIH Berlin Institute of Health @Charité