

ARX: Datenanonymisierung in Theorie und Praxis

Prof. Dr. Fabian Prasser

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

March 2022

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

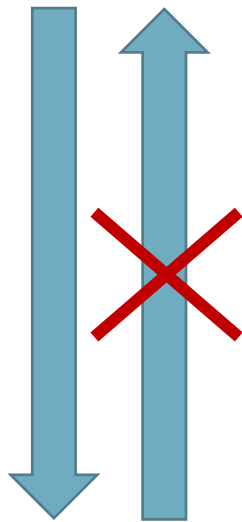
Initiativen zur Verbesserung der Transparenz, Reproduzierbarkeit und Wiederverwendbarkeit von Forschungsergebnissen und Forschungsdaten

- 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

Personenbezogene
Daten



Anonyme
Daten

DSGVO, Erwägungsgrund 26:

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

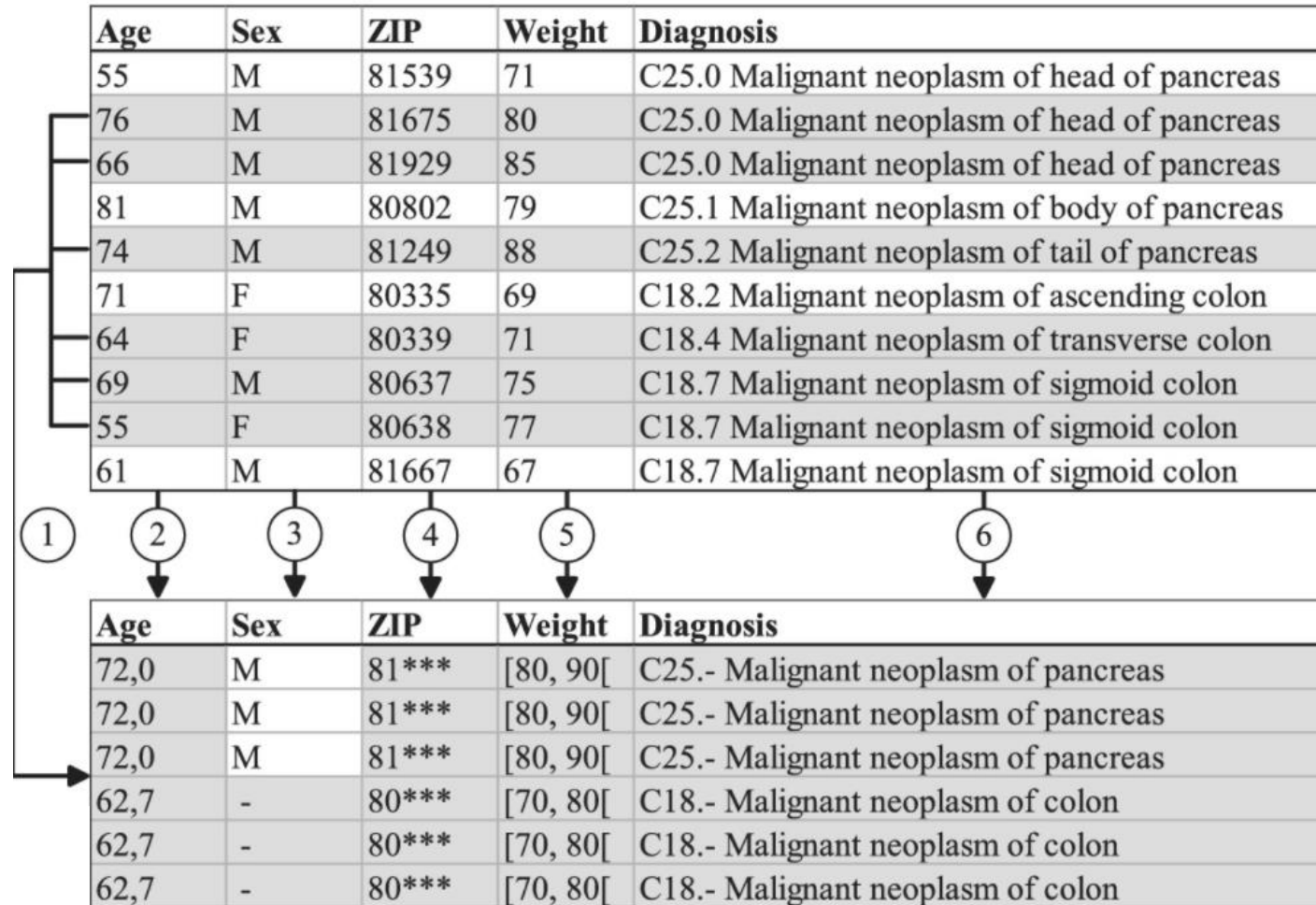
"[...] Um festzustellen, ob eine natürliche Person bestimmbar ist, sollten alle **Mittel berücksichtigt werden, die nach vernünftigem Ermessen eingesetzt werden können**, [...] um die natürliche Person direkt oder indirekt zu identifizieren [...]"

"[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)

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

Anonymisierung: Beispiel

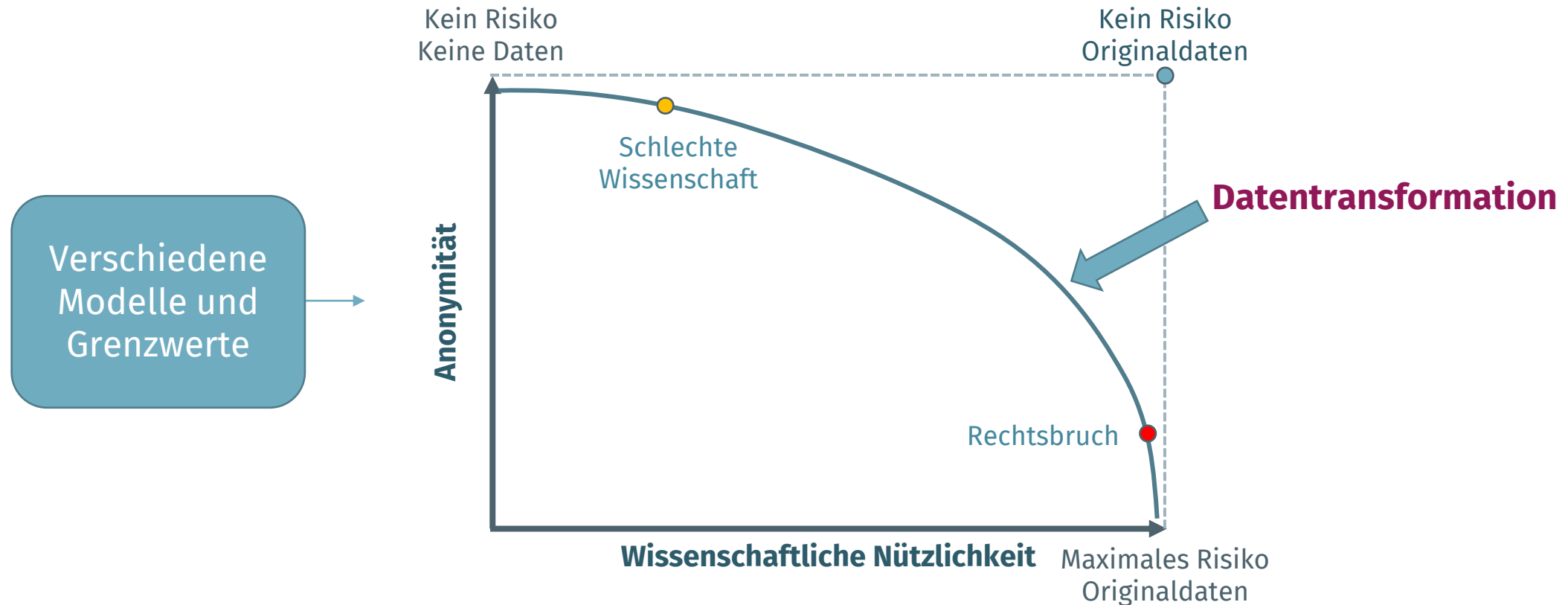


1. Stichprobenziehung
2. Aggregation
3. Löschung
4. Maskierung
5. Kategorisierung
6. Generalisierung

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.

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

Anonyme Daten: technische Perspektive



Source: Barth-Jones, Brussels Privacy Symposium, 2016

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

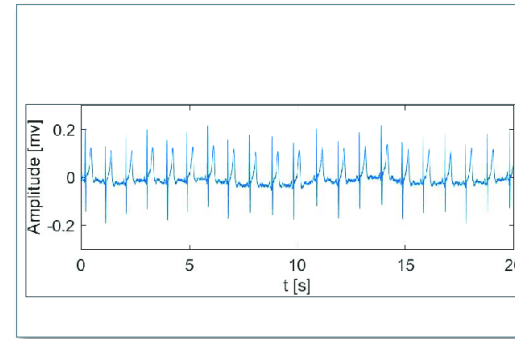
Risiken sind kontextabhängig!



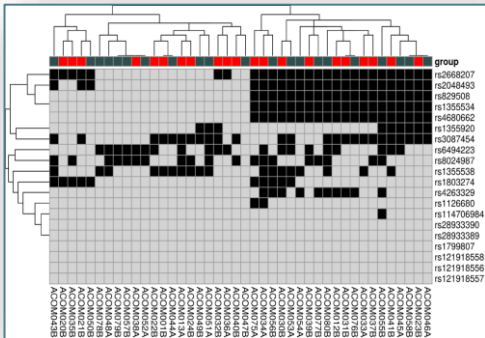
Source: https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface

AUTOPSY REPORT - Final Anatomic Diagnosis
 Dx: Sickle cell anemia with multiple red blood cell transfusions
 Cause of death per autopsy report (00-01-23): 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 8. From the age of 8 to 11/7/77, he had several health complications and underwent a liver transplant at the Case106 Hospital Center 14/3/November 2011. He has been in good health and continued with normal daily activities until Dec. 2055, 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, Mercedes Hesse. A few months later he began to experience chronic pain in his left hip and was referred to Dr. Soethe 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 5:00pm.

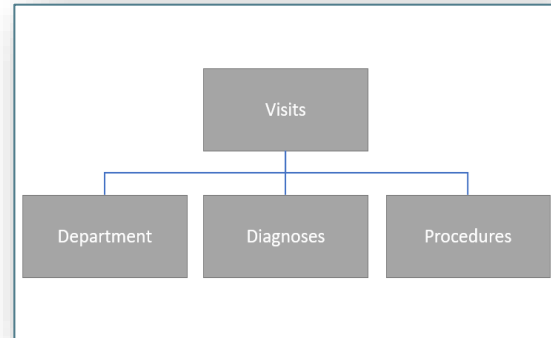
Source : <https://scrubber.nlm.nih.gov/>



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>

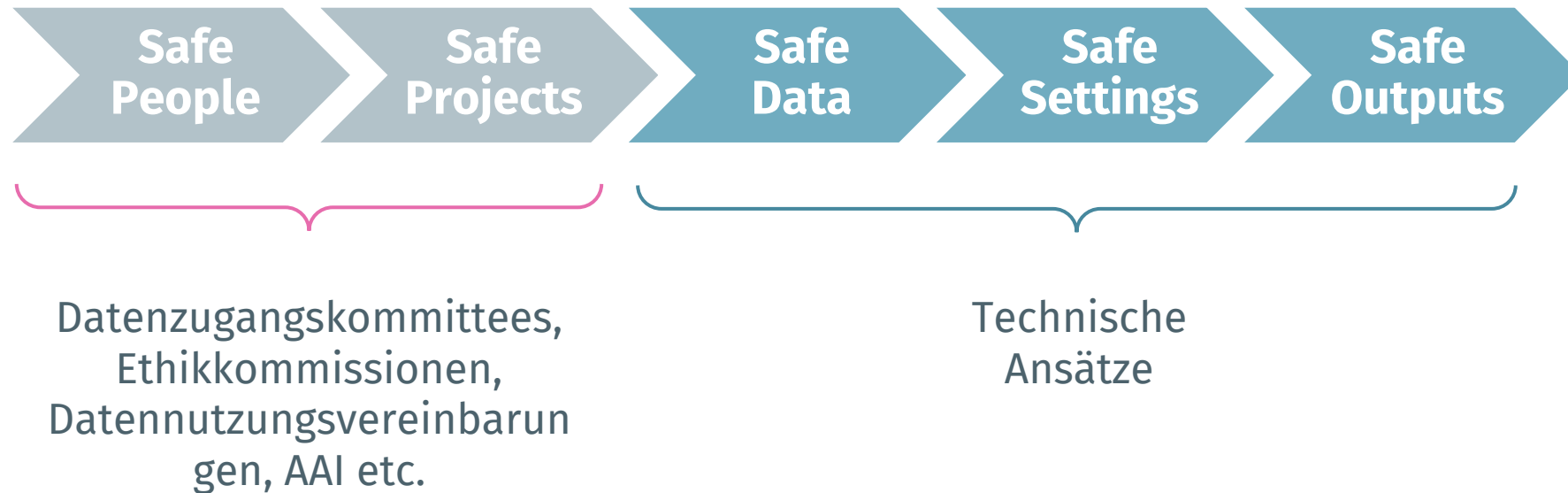
Marketer

Journalist

Prosecutor

Kontrolle des Kontexts

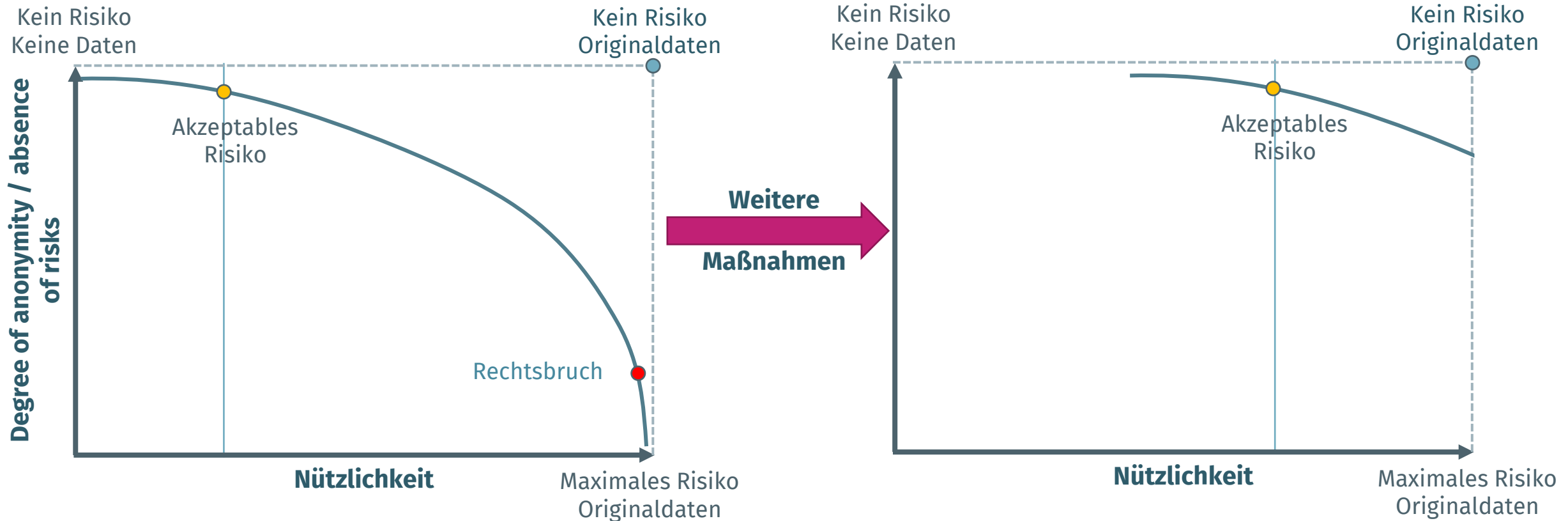
Das Five Safes Framework



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

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

Bedeutung weiterer Maßnahmen



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

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

ARX Data Anonymization Tool

The image displays four screenshots of the ARX Data Anonymization Tool interface, arranged around a central circular diagram. The central diagram consists of four quadrants: Configuration (red), Exploration (green), Risk analysis (yellow), and Quality analysis (blue), connected by arrows in a clockwise cycle.

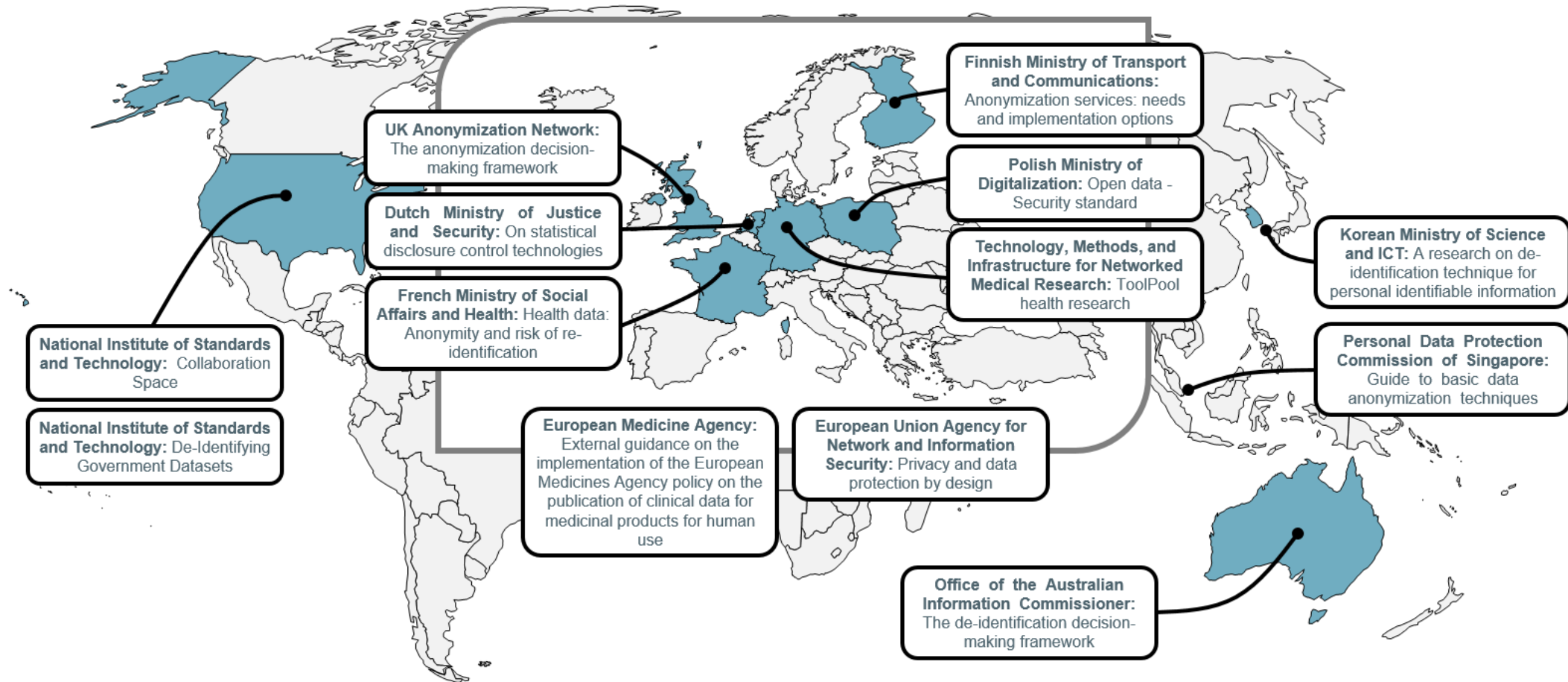
- Top-left screenshot:** Shows the 'Configuration' window. It displays a list of input data with columns for sex, age, race, marital-status, and education. A 'Transformation' window is open, showing a 'Generalization' rule for the 'education' attribute with levels 0-4.
- Top-right screenshot:** Shows the 'Exploration' window. It displays a grid of data points with various attributes highlighted in green and yellow. A 'Properties' window is open, showing details for a selected transformation, including 'Anonymity' and 'Minimal information loss'.
- Bottom-left screenshot:** Shows the 'Quality analysis' window. It displays a 'Summary statistics' window with a heatmap showing the distribution of data points across different categories.
- Bottom-right screenshot:** Shows the 'Risk analysis' window. It displays a 'Distribution of risk' histogram and a 'Re-identification risk' table. The table includes columns for 'Measure', 'Value [%]', and 'Population'. The 'Average prosecutor risk' is 80.340837%.

<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.

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

ARX: Impact



World Map provided by simplemaps.com

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.

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

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.

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

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

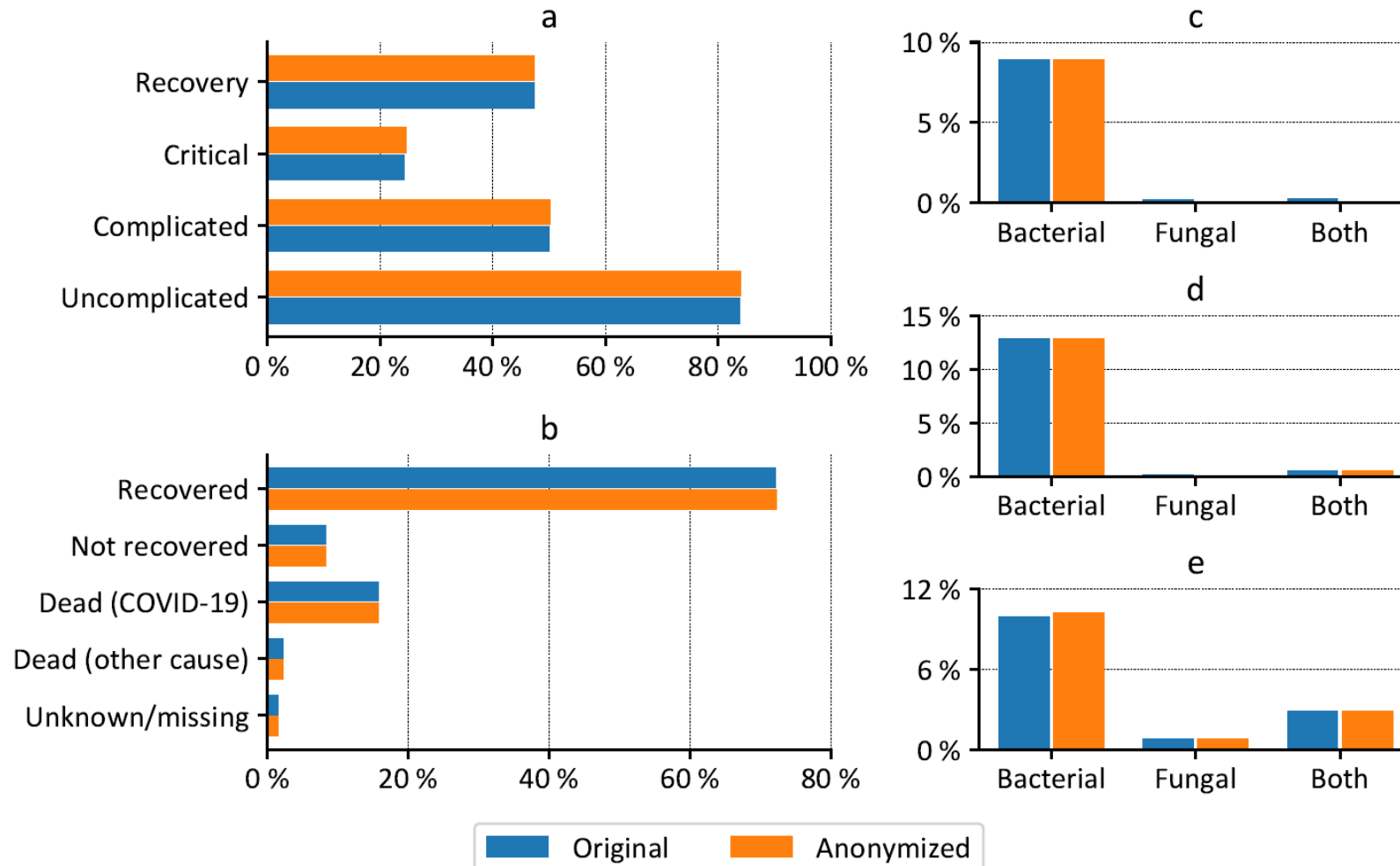
Auf der Grundlage der Stellungnahme zu Anonymisierungsmethoden der Artikel-29-Datenschutzgruppe (heute: Europäischer Datenschutzausschuss):

- 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.

Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

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. Online Workshop Open Data, Data Sharing und Datenschutz in der Medizin: Wunsch und Wirklichkeit

Danke für Ihre Aufmerksamkeit!

mi.bihealth.org

BIH Berlin Institute
of Health
@Charité