Herausforderungen bei der Anonymisierung (medizinischer) Forschungsdaten

DATENTAG: ANONYMISIEREN VON DATEN
07. Dezember 2022

Prof. Dr. Fabian Prasser

BIH Berlin Institute of Health @Charité

Motivation: Sekundärnutzung und „Data Sharing“

Innovationen: Datengetriebene Ansätze in der medizinischen Forschung

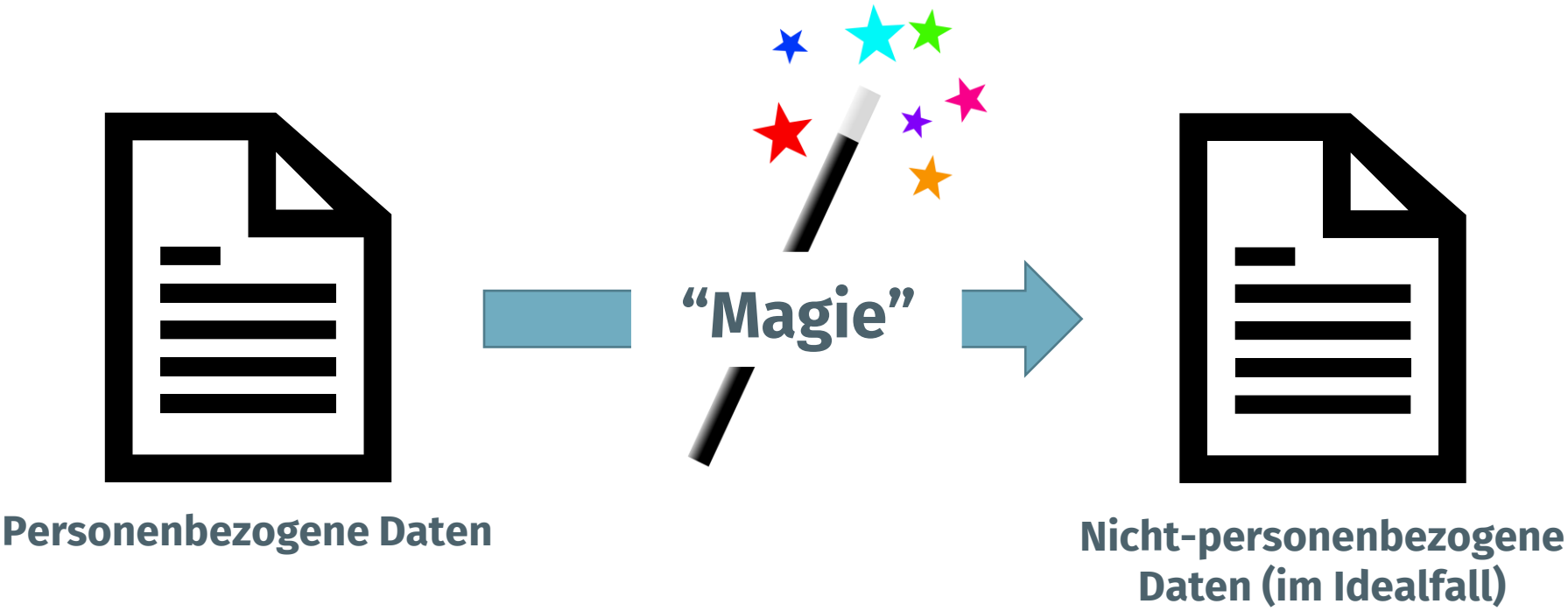
- Präzisionsmedizin: hohe Fallzahlen, detaillierte Charakterisierungen
- Real-World Evidence: Sekundärnutzung, z.B. von klinischen Routinedaten für die Forschung
- Kollaborative Forschung, z.B. gemeinsame Nutzung von Daten über institutionelle Grenzen hinweg

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

- 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

→ Anonymisierung: Datenminimierung, gesetzliche Vorgaben (vgl. ZfKD, FDZ), Herausforderungen bei einwilligungsbasiertem Vorgehen, mangelnde Rechtsgrundlage

Grundkonzept der Anonymisierung



Anonymisierung von tabellarischen Daten

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

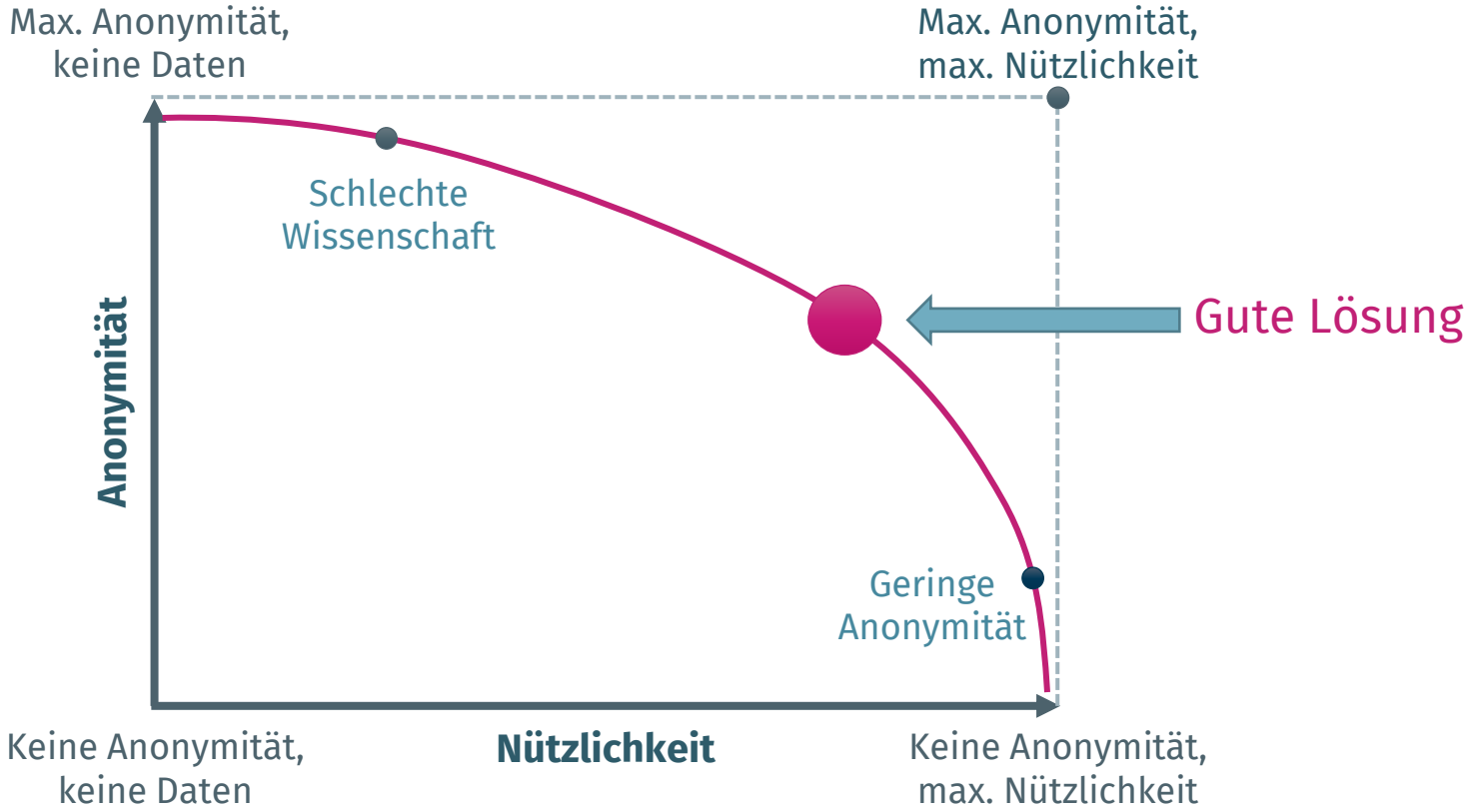
Alter	Geschlecht	PLZ	Gewicht	Diagnose
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
72,0	Männlich	81***	[80, 90[	C25.- Bösartige Neubildung des Pankreas
62,7	---	80***	[70, 80[	C18.- Bösartige Neubildung des Kolons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

- Sampling
- Aggregation
- Entfernen
- Maskierung
- Kategorisierung
- Generalisierung

→ k-Anonymity (mit k=3) und (ε, δ)-Differential Privacy (mit ε ≈ 0.92 und δ ≈ 0.22)

Herausforderung 1: Restrisiken akzeptieren

Personenbezogene Daten?
Re-Identifizierung?
Akzeptable Restrisiken?



Vorgesehene Nutzung?
Weitere Anforderungen?

Herausforderung 2: Selektion von schützenswerten Variablen

Katalog-basiertes Vorgehen

- Vergleich mit in Gesetzen und Leitfäden als "riskant" eingestuften Variablen

Qualitative Risikoanalyse nach Malin et al.

Replizierbarkeit: Wahrscheinlichkeit, dass Merkmale in Bezug auf die Betroffenen immer wieder auftreten	**Gering**: Blutzuckertests variieren **Hoch**: Demografische Daten sind recht statisch
Verfügbarkeit: Welche externen Ressourcen können replizierbare Merkmale enthalten und wer hat zu diesen Zugang?	**Gering**: Laborbefunde sind meist nur im Gesundheitsbereich bekannt **Hoch**: Demografische Daten sind z.B. in Sterbe- und Heiratsregister enthalten.
Unterscheidbarkeit: Ausmaß, in dem Merkmale Probanden unterscheidbar machen	**Gering**: Geschlecht **Hoch**: Seltene Erkrankung

- Gefolgt bspw. von einer Schwellwertanalyse

Quantitative Verfahren: Uniqueness, Separation

Herausforderung 3: Messung und Reduktion von Risiken

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

- **Aussonderung**: Möglichkeit, einige oder alle Datensätze zu isolieren, die eine Person im Datensatz beschreiben
  - K-Anonymity, statistische Modelle, Minimum Sample Uniques, ..
- **Verknüpfbarkeit**: Möglichkeit, mindestens zwei Datensätze zu verknüpfen, die dieselbe Person oder eine Gruppe von betroffenen Personen betreffen
  - K-Anonymity, Minimum Sample Uniques, ...
- **Inferenz**: Möglichkeit, mit erheblicher Wahrscheinlichkeit den Wert eines Attributs aus den Werten einer Reihe anderer Attribute abzuleiten.
  - t-Closeness, ℓ-Diversity, β-Likeness, ...

NHS Digital: ISB1523 - Anonymisation Standard for Publishing Health and Social Care Data

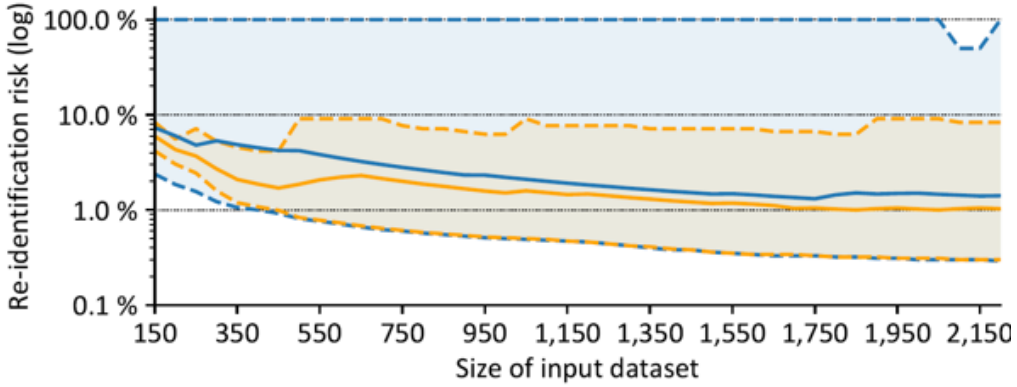
- Eine freie Variable, k-Anonymity mit k abhängig vom Datensatz

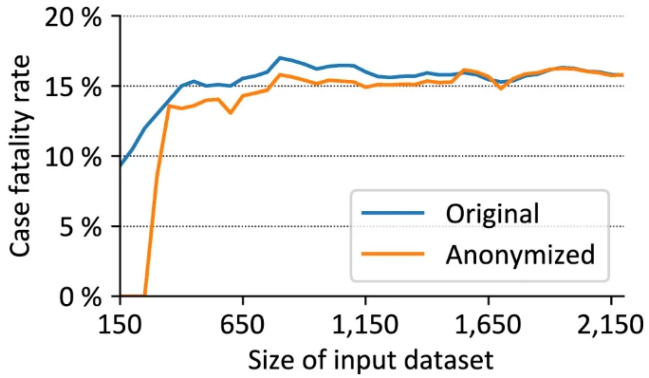
Herausforderung 4: Umgang mit hochdimensionalen Daten

Kombination von strikten, formalen Methoden mit „Faustregeln“

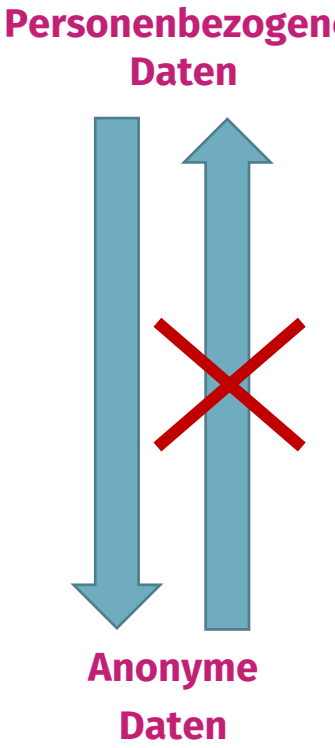
Beispiel: LEOSS-Register

- Quantitative und qualitative Risikoanalyse
- Formale Methoden für besonders risikoreiche Variablen
- Weitere Maßnahmen für weitere Variablen
  - Ausgewählte metrische Variablen kategorisieren
  - Zeitangaben relativieren und verschieben
  - Gruppierung oder Entfernung sensibler Variablen
- Kompensation von Restrisiken durch abgestuftes Modell
  - Public Use File
  - Scientific Use File (Datennutzungsvertrag, Identitätsprüfung)





Herausforderung 5: Anonymität als Prozesseigenschaft (1)



DSGVO, Erwägungsgrund 26:

"Die Grundsätze des Datenschutzes sollten für **alle Informationen gelten, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen** [...]"

"[...] Um festzustellen, ob eine natürliche Person identifizierbar ist, sollten alle **Mittel berücksichtigt werden, die** [...] **nach allgemeinem Ermessen wahrscheinlich genutzt werden,,** [...] um die natürliche Person direkt oder indirekt zu identifizieren [...]"

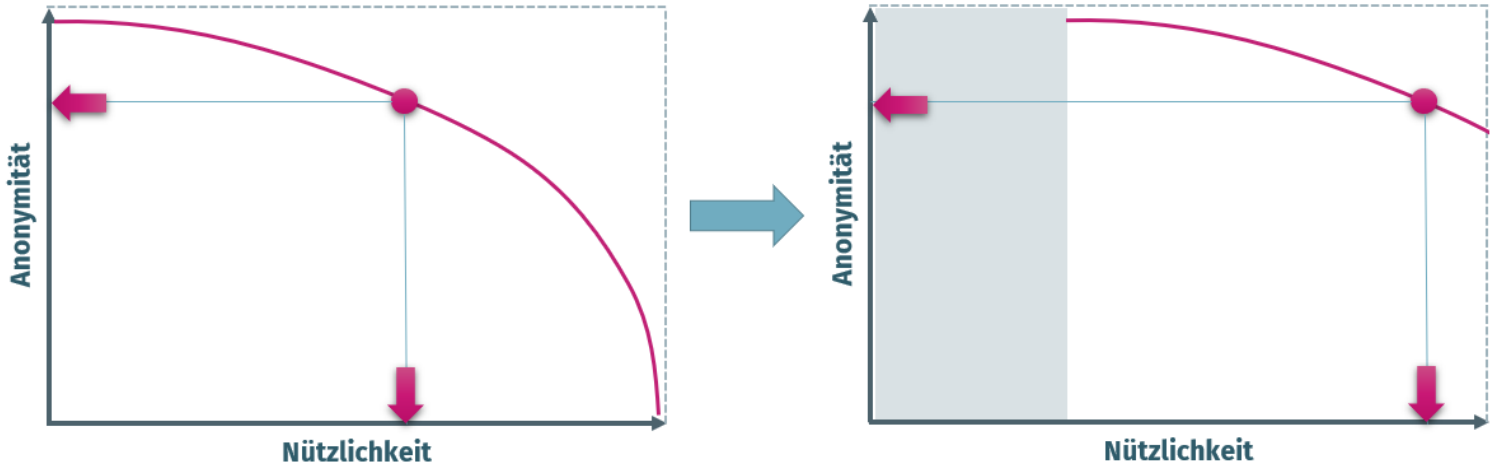
"[Dabei] alle **objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand**, herangezogen werden, wobei die zum Zeitpunkt der Verarbeitung **verfügbare Technologie und technologische Entwicklungen** zu berücksichtigen sind. [...]"

Herausforderung 5: Anonymität als Prozesseigenschaft (2)

Anonymität nicht als reine Dateneigenschaft...

- ...sondern als Eigenschaft eines Verarbeitungsprozesses, der weitere Maßnahmen enthalten kann

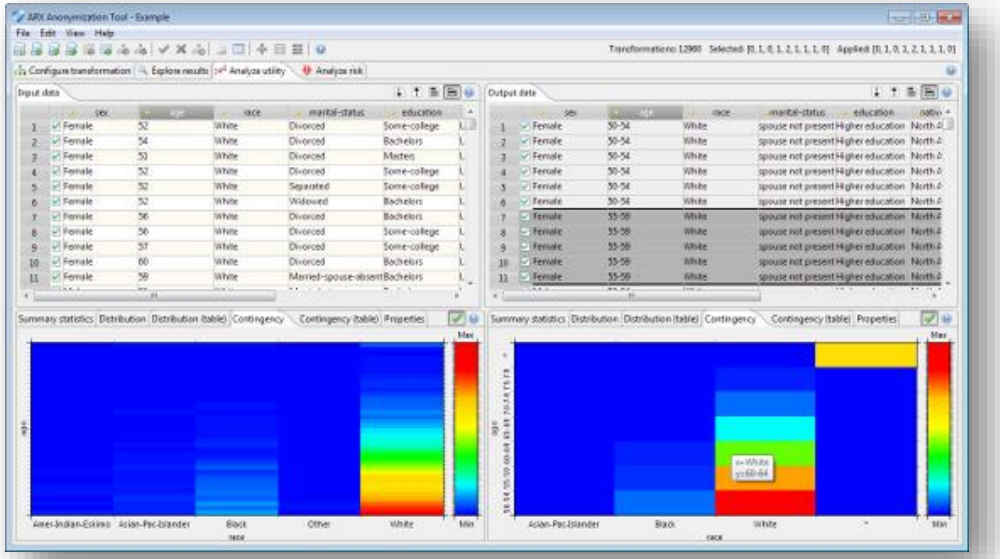
Nur dann ist Anonymisierung auch für komplexe Daten und Nutzungsszenarien praktikabel



- **Vergleiche auch Differential Privacy**

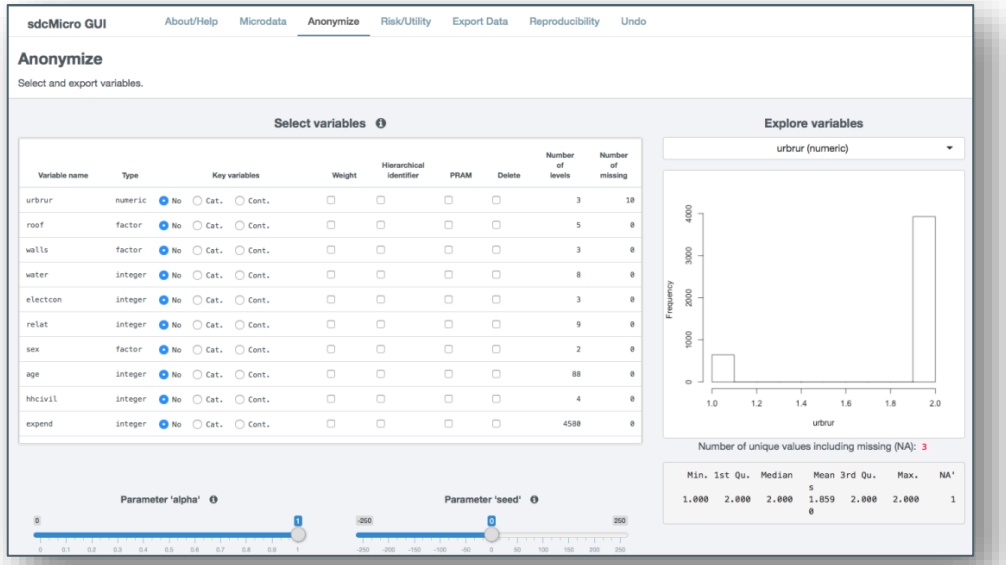
Herausforderung 6: Werkzeugunterstützung

- **Nur wenige frei verfügbare Werkzeuge sind praxistauglich**



ARX Data Anonymization Tool

sdcMicro



- **Implementierung von Anonymisierungswerkzeugen ist herausfordernd**

Danke für Ihre Aufmerksamkeit!

Prof. Dr. Fabian Prasser
Medizininformatik
Berlin Institute of Health @ Charité

https://mi.bihealth.org
fabian.prasser@charite.de

BIH Berlin Institute of Health @Charité