EUROPE BIOBANK WEEK

SEPTEMBER 13-16, 2016
VIENNA - AUSTRIA

Fabian Prasser: "Sharing anonymous biomedical data: tools, trade-offs and perspectives"

BBMRI-ERIC® gateway for health

BBMRI.at Biobanking and BioMolecular resources Research Infrastructure Austria

European, Middle Eastern & African Society for Biopreservation & Biobanking

Sharing anonymous biomedical data: trade-offs, tools and perspectives

Fabian Prasser

Chair for Medical Informatics (Prof. Klaus A. Kuhn)
Institute for Medical Statistics and Epidemiologie

University Hospital rechts der Isar
Technical University of Munich (TUM)
Germany

Motivation and background

Motivation: legal requirements and acceptance

- Secondary use of data, e.g. clinical data for annotating biospecimens
- Sharing of data in cooperative research, e.g. networked biobanking
- Maintaining societal and individual acceptance, e.g. of donors

Goal: privacy protection

- Make it very difficult for recipients to learn the identity of data subjects
- Make it very difficult for recipients to associate individuals with sensitive information

Process: data anonymization

- **Step 1:** Remove directly / obviously identifying information (e.g. names, insurance numbers)
- **Step 2:** Modify data to reduce the uniqueness of potentially identifying attribute values (e.g. date-of-birth, sex, zip code), which may in combination uniquely identify patients or probands within a given context

Example: data anonymization

sex	age	race	marital-status	education	native-coun...
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		

Reduction of the uniqueness of potentially identifying values

sex	age	race	marital-status	education	native-coun...
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		Mexico
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		
Female		White	Married-civ-spouse		

Generalization

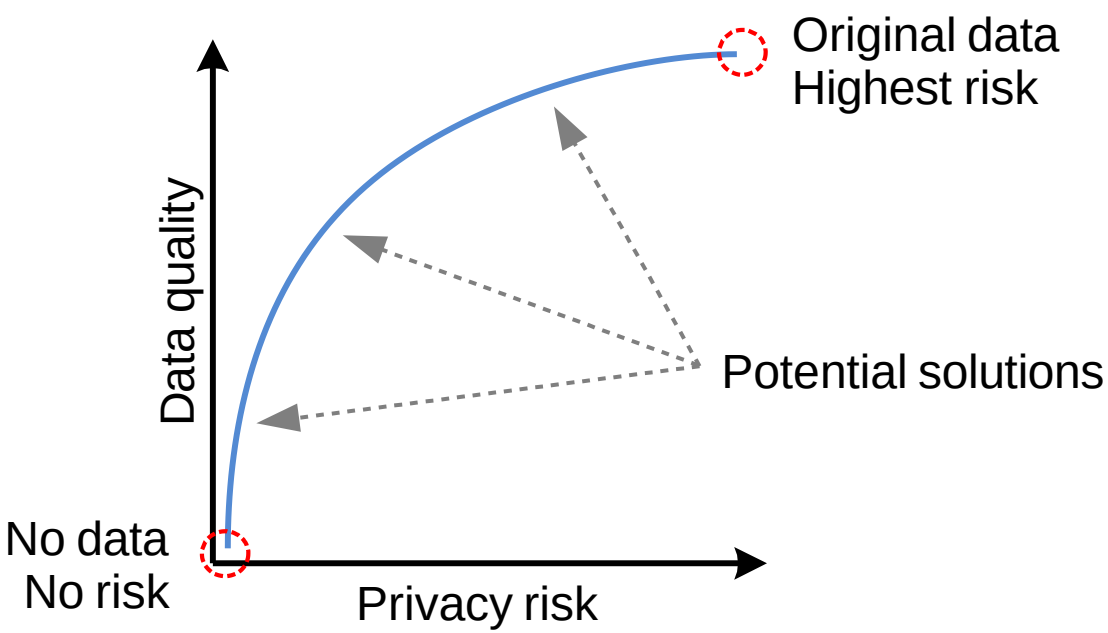
Suppression

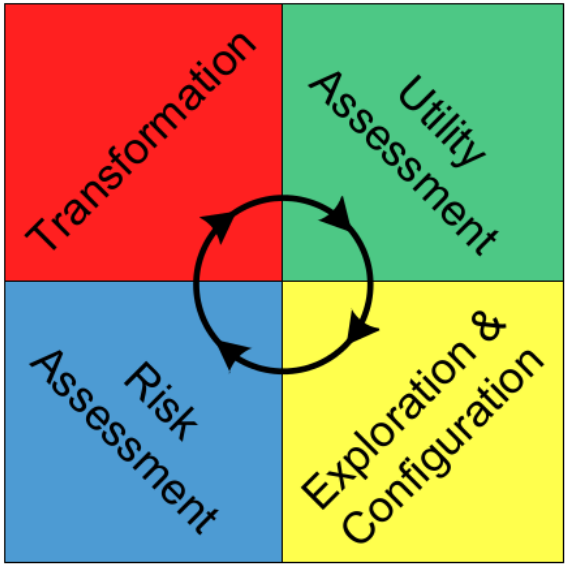
Micro-aggregation

Trade-offs

Central trade-off: privacy risks vs. usefulness of data

- **Usefulness:** flexibility of data processing *or* quality of data



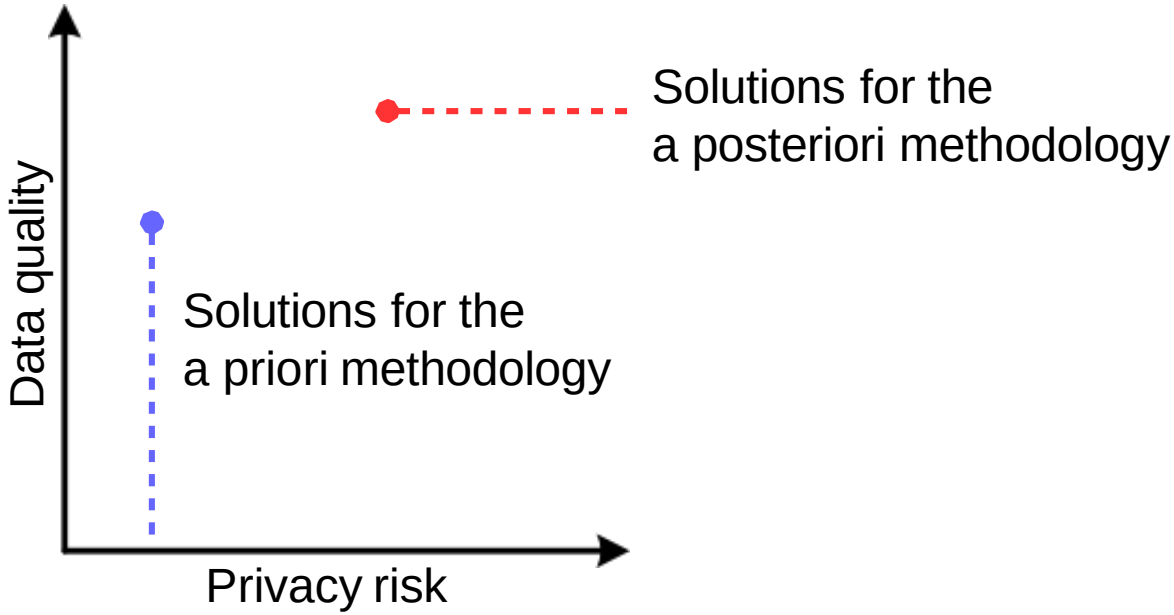


Models and methods are needed for analyzing and quantifying both aspects

Features of anonymization tools

Basic methodology

- **A priori risk control:** quality follows risk (data anonymization)
- **A posteriori risk control:** risk follows quality (statistical confidentiality)



Transformation methods

- **Truthful, e.g.:** attribute generalization, record suppression
- **Pertubative, e.g.:** microaggregation, rank swapping, noise addition

Features of anonymization tools

Ability to handle different types of input data

- **Relational data:** tabular data
- **Longitudinal / transactional data:** repeated collection of the same information
- **Data with relational and transactional characteristics**
- **Scales of measure and data types:** categorical, numeric, continous, dates, …
- **Dimensionality of data:** high-dimensional vs. low-dimensional data
- **Scalability:** large vs. small datasets
- **Data with clusters:** e.g.: household structures

Models for quantifying privacy risks

- **For example:** k-anonymity, k-map, strict average risk, population uniqueness …

Methods for analyzing and models for quantifying data quality

- **For example:** loss of information (e.g. granularity), changes in statistical properties (e.g. tendency, dispersion, shape of distributions), data utility (e.g. statistical classification)

Mature open source tools: μ-ARGUS [ARG]

Developed at Statistics Netherlands

- Under development since 1990

Primary application domain

- Disclosure control in official statistics

Methods

- Focus on the a posteriori methodology

Interfaces

- Cross-plattform graphical user interface
- No programming library

Scalability

- Up to a hand full of potentially identifying attributes
- Small to medium sized data

Mature open source tools: sdcMicro [SMG]

Developed at Statistics Austria / Technische Universität Wien

- Under development since 2008

Primary application domain

- Disclosure control in official statistics

Methods

- Focus on the a posteriori methodology
- Highly flexible because of R integration

Interfaces

- Primarily a package for the R statistics software
- Cross-plattform graphical user interface

Scalability

- Highly scalable when used as a programming library
- Scalability issues when using the graphical interface

Mature open source tools: ARX [ARX]

Developed at Technical University of Munich

- Under development since 2011

Primary application domain

- Anonymization of biomedical data for research purposes

Methods

- Focus on the a priori methodology
- Focus on truthful transformation methods
- Supports a wide variety of approaches to data de-identification

Interfaces

- Comprehensive cross-plattform graphical user interface
- All functionality also available as a Java programming library

Scalablity

- Millions of records with up to 50 potentially identifying attributes

Further tools

Research prototypes: lack of scalability and functionality

- **UTD Anonymization Toolbox** [UTD]
- **Cornell Anonymization Toolkit** [CAT]
- **OpenAnonymizer** [OAN]
- **TMF AnonTool** [TAT]

Commercial and closed-source software

- **SECRETA** [SEC]
  - Research prototype that can handle transactional data
- **Privacy Analytics CORE** [PAC]
  - Leading commercial software for health data de-identification in US & Canada
  - Comparable to open source solutions but offers „enterprise“ features

Further tools: little information available

- **IHSN tool** [IHSN]
- **CATS platform** [CCP]

Perspectives: primary and secondary measures

Privacy risks can never be reduced to completely zero and data quality is of utmost importance

- Secondary measures must be implemented to control residual risks and to ensure that the required degree of data quality can be preserved

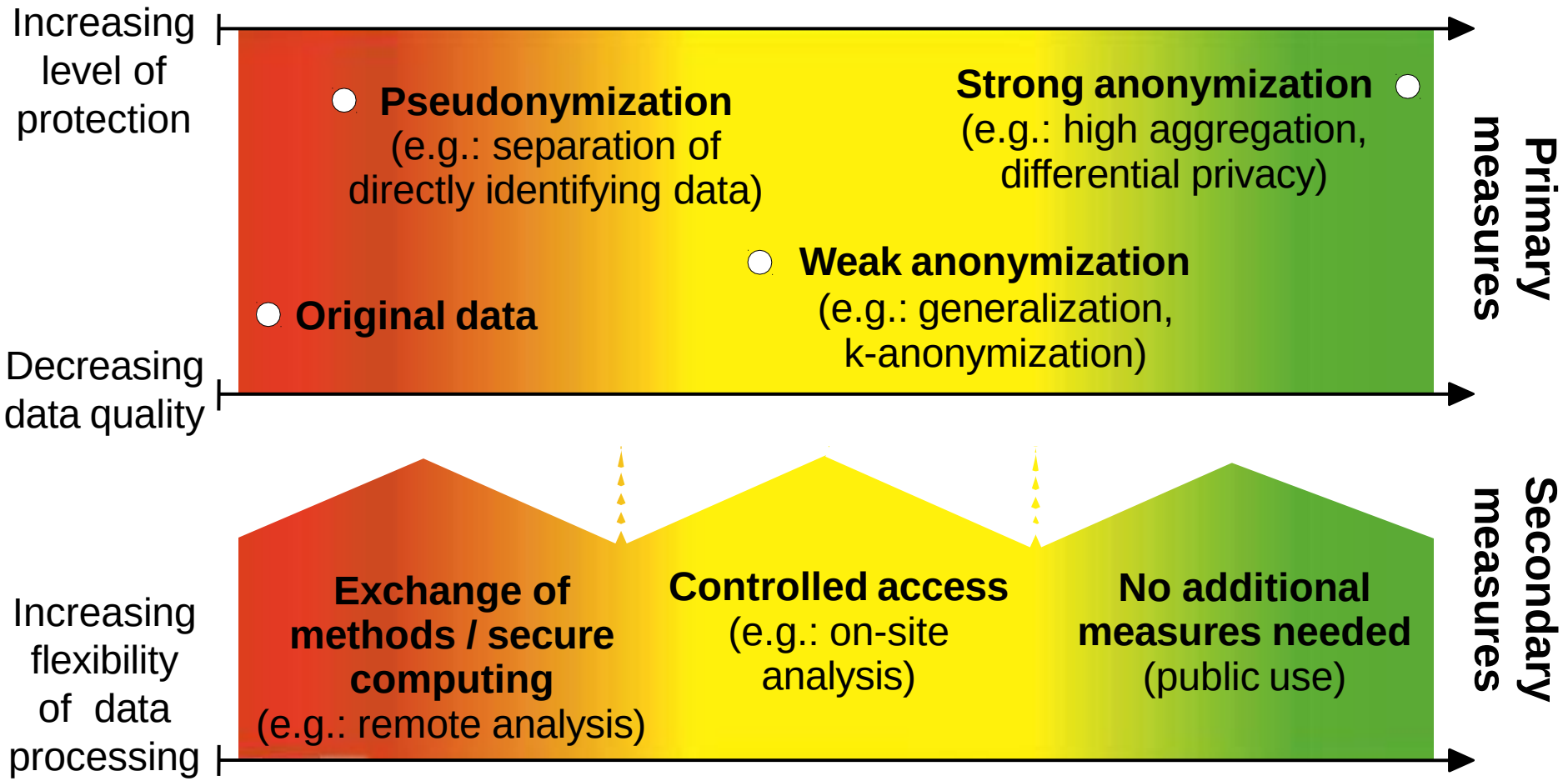
Make sure that the recipient is as trustworthy as possible

- Sign data use agreements
- Install data access committees
- Implement multiple levels of access

Secondary technical measures must be chosen depending on context

- Which organizational and legal measures have been installed?
- How sensitive is the data?
- How trustworthy is the recipient?
- How many individuals would be affected by a privacy breach?

Perspectives: secondary technical measures



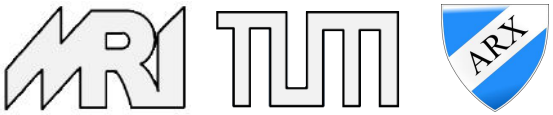
Thank you for your attention!

Dr. rer. nat. Fabian Prasser

Klinikum rechts der Isar
Technische Universität München
Institut für Medizinische
Statistik und Epidemiologie

Ismaninger Str. 22
81675 Munich
Germany

Tel +49 89 4140-4328
Fax +49 89 4140-4850
fabian.prasser@tum.de
www.imse.med.tum.de
arx.deidentifier.org



References

[ARG] http://neon.vb.cbs.nl/casc/mu.htm

[ARX] http://arx.deidentifier.org

[CAT] http://sourceforge.net/projects/anony-toolkit/files/

[CCP] https://www.custodix.com/index.php/cats

[IHSN] http://www.ihsn.org/home/projects/sdc

[OAN] http://sourceforge.net/p/openanonymizer/

[PAC] http://www.privacy-analytics.com/software/privacy-analytics-core/

[SEC] http://users.uop.gr/~poulis/SECRETA/

[SMG] https://cran.r-project.org/package=sdcMicro

[TAT] http://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx

[UDT] http://cs.utdallas.edu/dspl/cgi-bin/toolbox/