SCTO PLATFORMS

Facts and pitfalls of observational studies

Data Governance and Protection: Practical Experience

Constantin Sluka, PhD

HRO Lunch Seminar, 17.04.2024

Introduction

HRO

Further Use of Data (Chapter 3)

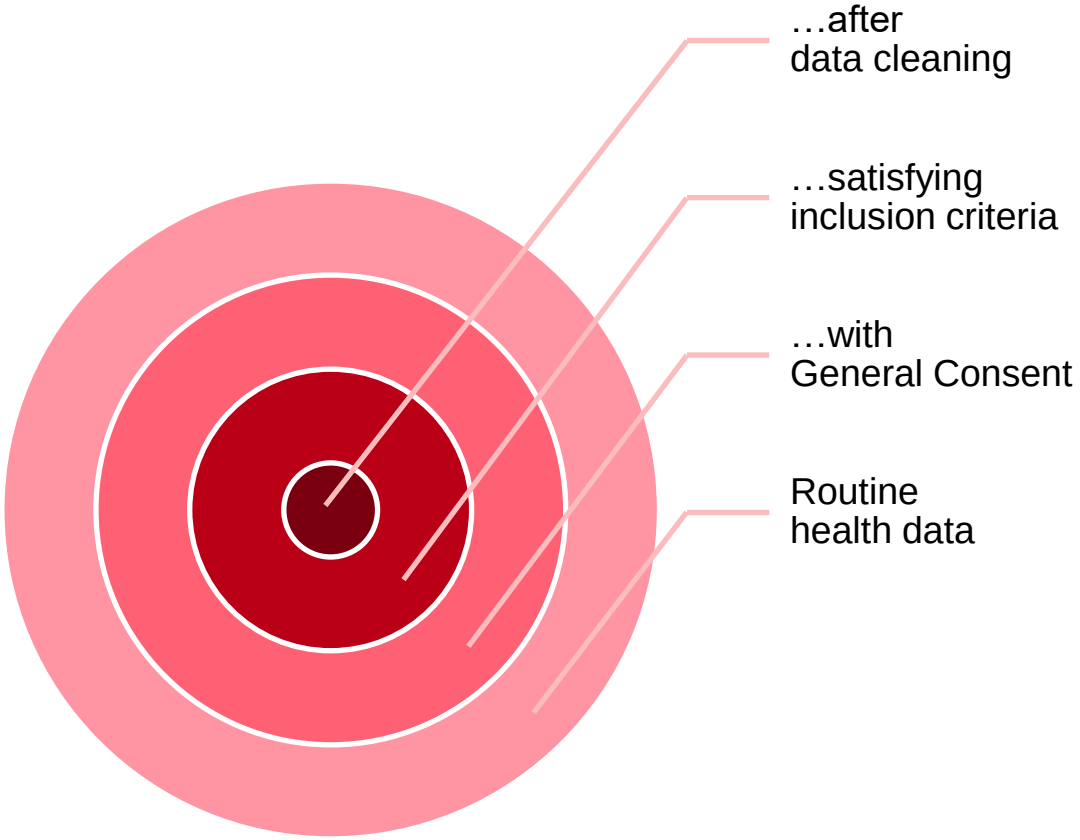
- Health data already collected (during routine)

Research Projects with Persons (Chapter 2)

- Data collection to
  - Answer a scientific question
  - Make further use of data later on

SCTO PLATFORMS

How to get routine data for research?



- Health data is routinely collected in clinical (primary) systems
- Only data from patients with signed (general) consent may be further used for HRO studies
- Before start of your project: Assess data availability and data quality

SCTO PLATFORMS

How to get routine data for research?

Feasability Request:

- How many patients match the criteria for your project?
- How complete/clean are the data?
- Needs to be done before start of your project
- No need for ethics approval if only aggregated numbers are returned (*"n = 100 patients satisfy your criteria"*)

SCTO PLATFORMS

How to get routine data for research?

Pitfalls:

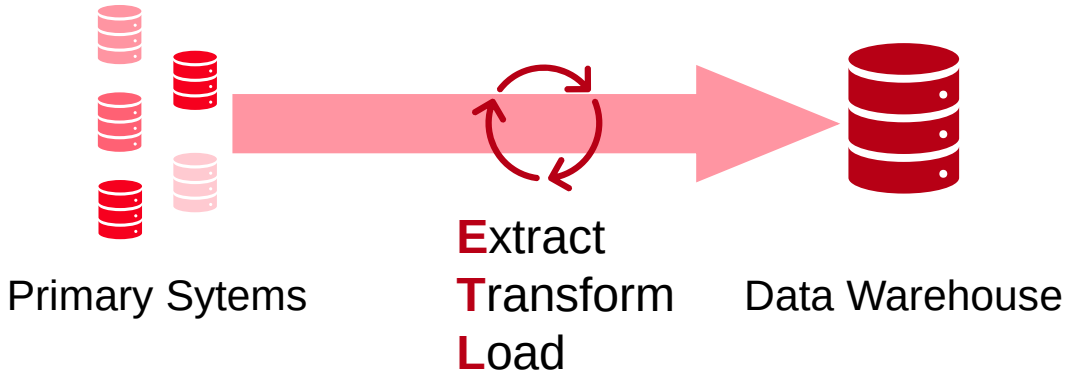
- Routinely collected health data **are not** collected for research!
- Lack of structure: Abundance of free text results (e.g. diagnoses)
- Lack of interoperability: Each application uses unique data model
- Lack of standards: Multiple different variables with same information
- Lack of continuity: Constant change of processes and systems

Recommendations:

- Check if the data you need can be extracted
- You will have to invest in data cleaning
- You will have to understand who collected what data for what purpose
- Plan and budget these steps before submitting your project

How to get routine data for research?

Clinical Data Warehouse (CDWH):



- Integrates data from multiple primary systems into a single data model.
- Simplifies **F**indability and **A**vailability of data (2 FAIR criteria).
- No (or little) data cleaning (domain knowledge needed).

SPHN:

- Supported University Hospitals in the setup of CDWHs
- Defines terminologies for some standardized variables («Meta-data Catalogue»
- Federated Query System allows feasability requests across all five University Hospitals

sphn.ch/network/projects/collaboration-agreements-with-hospitals/

SCTO PLATFORMS

How to get routine data for research?

Feasability Request

- Number of patients with General Consent
- Availability and structure of data

Planning

- Budget for Data Management and Data Cleaning
- Writing of research plan and proposal
- Submitting of project to Ethics Committee(s) and **Data Governance Board(s)**

Data Extraction

- Signature of Data Transfer and Usage Agreement (DTUA)
- Pseudonymisation or Anonymisation
- Secure transfer of data

Research

- Data cleaning and post-processing
- Structuring data with help of domain knowledge
- Data analysis

SCTO PLATFORMS

Routine data vs prospective data collection

Routine data		Prospective data collection	
✔	Data **existing**	**Only future data** available	✘
✔	Large N and Long time periods	Patient years grow slowly	✘
✔	Data model **existing**	**Specification** of data model needed	✘
✘	**Unstructured** data	Data **structured**	✔
✘	Data **not** collected for research	Data collected for **your research**	✔
✘	Only variables from **clinical routine**	Exactly the **variables you specify**	✔
✔	No additional manual data entry	Error prone manual data entry	✘
✘	No additional manual data entry	Consistency check during data entry	✔
✘	**Data cleaning** after extraction	**Data cleaning** during collection	✘

Of course, you can prospectively collect a few new variables AND make further use of routine data. → Pitfall: Matching of patients and visits in CDWH

SCTO PLATFORMS

Data storage and access to data

Data protection by design and by default:

- Your organisation and IT department should ensure organisational and technical measures for data protection.
- You should use secure and encrypted IT infrastructure maintained by your IT department.
- Make sure data are backed-up and protected against accidental alteration.
- **Data Minimisation:** Only collect data needed for your specific project.
- Use pseudonymised data, Anonymise if possible.
- **Never** share unencrypted sensitive data via e-mail / messenger / cloud!
- **Never** share your passwords!

SCTO PLATFORMS

Reproducible Data Analysis

Create reproducible data workflows:

- Strictly separate the analysis from the data, never write over raw data!
- Do not use spreadsheet applications:
  - Auto correction changes data
  - Type and format of variables are open
  - Not reproducible, not GCP compliant
- Use electronic data capture software for prospective data collection
  - User authentification with password
  - Controlled data format
  - Audit Trail

SCIENCE / TECH / MICROSOFT

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

/ Sometimes it's easier to rewrite genetics than update Excel

NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

SCTO PLATFORMS

Pseudonymisation vs Anonymisation

Pseudonymisation (Coding)

- Patients are **coded** with a (patient) ID
- The key is stored separately from the data
- All direct identificators are removed

Pseudonymised data remain identifieable personal data

Anonymisation

- Re-identification is only possible **with disproportionate effort**
- HRA and FADP are out of scope for anonymised data
- All identificators have to be removed or transformed

There is no rigid definition of anonymisation in EU & CH!
Assess and mitigate the risk of re-identification

SCTO PLATFORMS

Direct Identifiers

Can be directly used to uniquely identify a specific person

- Names, addresses, dates of birth, …
- Unique identification numbers: Social security number, (hospital) patient ID, serial numbers of devices, …
- Event dates: Admission, death, …
- Full-face photographs, fingerprints, voice recordings, genetic data, …

⇒ **Remove**

SCTO PLATFORMS

Indirect Identifiers

Can be linked to other data sources for re-identification

- Rare diseases, treatments, biomarkers, …
- Any high-precision variable (depends on type of data)
- Any uncommon value or combination of variables

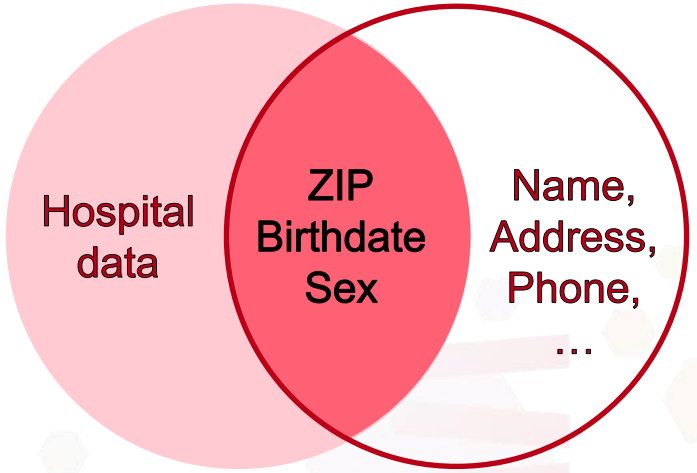
⇒ Transform to mitigate the risk of re-identification

SCTO PLATFORMS

Linkage-attacks

The process of re-identifying a person in a dataset by linking it to external data containing person identifying information.

Famous example from 1997:
Latanya Sweeney re-identifies the governor of Massachusetts in “anonymised” medical data of Massachusetts state employees by linking it with the Cambridge voter registration list.



L. Sweeney, **Simple Demographics Often Identify People Uniquely.** https://dataprivacylab.org/projects/identifiability/paper1.pdf
L. Sweeney, **Only You, Your Doctor, and Many Others May Know.** Technology Science. 2015092903. September 28, 2015. https://techscience.org/a/2015092903/

SCTO PLATFORMS

Anonymisation: A risk-based process

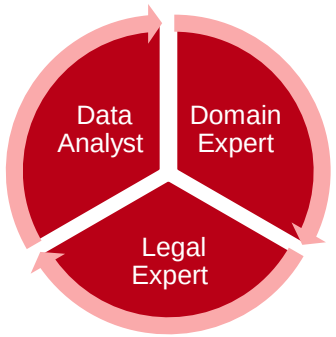
The risk of re-identification is a function of context risk and data risk

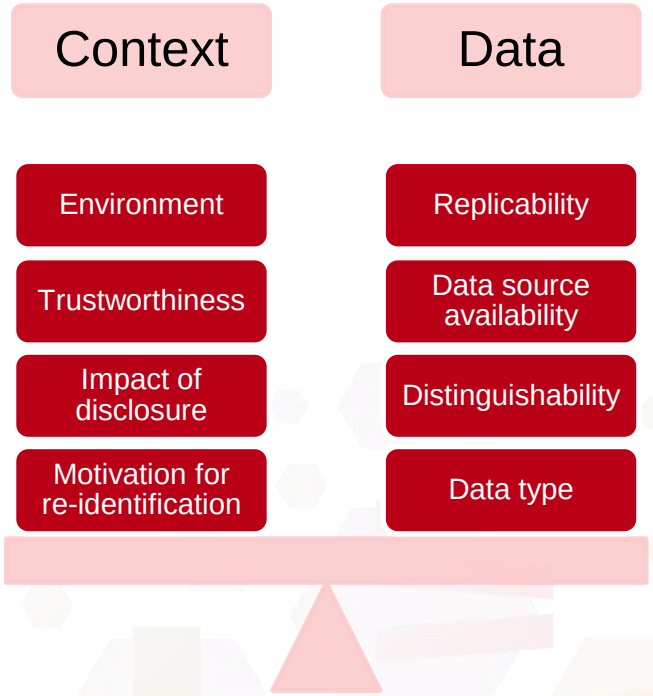
Context risk: The probability of an attack*

*or accidental disclosure or data breach

Data risk: The probability of identity disclosure given there is an attack

⇒ Mitigate disclosure risk below an acceptable threshold





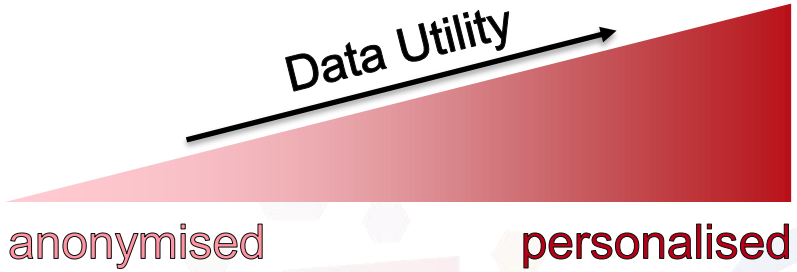
Open data vs Controlled Access: Strong control of environment is an efficient way to mitigate disclosure risk

SCTO PLATFORMS

Data Transformations and Information Loss

Data Transformations

- «Top and bottom» Recoding, Suppression, (Micro-) Aggregation, Rounding, Random noise, Permutations,...



Anonymisation has impact on data utility

- Data Transformations lead to information loss
- You need to check if anonymised data and information loss are suitable for your project
- You need to document all data transformations, information loss, and the remaining re-identification risk

SCTO PLATFORMS

Summary

Routine health data are collected for health care! Know the limitations and budget for data management and cleaning before starting a research project.

Clinical Data Warehouses simplify data extractions by integrating data from different clinical applications into a single data model.

Ethics Committees and the **Data Governance Boards** of the data controlling health care providers **need to approve** a HRO project with routine data.

Make sure to use secure, encrypted, and backed-up infrastructure maintained or endorsed by your IT department.

Use reproducible data workflows that separate analysis and raw data.

Pseudonymised data remains personal data. **Anonymisation is a risk-based process** and will lead to information loss.

SCTO PLATFORMS

Thank you

for your attention.