EL1: Intro

Overwiev of this part of the course

Important aspects

- Data (where does is come from, what does it contain)
- Ethics and legal (how to handle sensitive data, what laws and regulations apply)
- Project management (how to plan and execute a data project, version control, reproducibility, R specific packages for efficient data handling)

Data – what is it?

EU Data Act | Article 2, Definitions:

> For the purposes of *this Regulation*, the following definitions apply:
>
> (1) ‘data’ means any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audio-visual recording;
>
> (2) ‘metadata’ means a structured description of the contents or the use of data facilitating the discovery or use of that data;
>
> (3) ‘personal data’ means personal data as defined in Article 4, point (1), of Regulation (EU) 2016/679;
>
> (4) ‘non-personal data’ means data other than personal data;

Course structure

- Lectures on different data sources/registers
- 🧑‍💻 Exercises on data management and analysis
  - R with some additional tools (Git, GitHub, targets, data.table)
- A data project with written report and presentation
- Final exam 😱
- Instruction web page in addition to Canvas
- Litterature: accassible through GU library (O’Reilly Learning for Higher Education) or otherwise shared (no need to purchase books)

Different types of data

- 🩻 Images
  - Statistical image analysis
- 🧪 Lab samples
- 📝 Unstructured medical records
  - Natural Language Processing
- 📺 Sensor data
  - Time series (“big data”)
- 💻 EHRs (electronic health records)
  - Structured but hierarchical rather than tabular
- 💻 **Structured medical records**
  - tabular data

Usages

- 👩‍🔬 Research
- 📒 Quality control/improvement
- 🏷️ Administration/reporting
- 📰 News coverage
- 🧑‍💻 Building prediction models and tools

Register data

Three types of health care registers:

- Administrative registers
- Health care registers
- Quality registers

💶 Administrative data

(As found in all types of registers)

- Billing codes
  - Direct (what something actually cost)
  - Estimated (DRG codes for different types of procedures)
- Claims data
  - Primary for reimbursment (insurense company or other payer)
  - Secondarily for Health economy/epidemiology
- How to contact patients, health care providers etc
- Dates and times for visits, procedures etc

🏥 Hospital background data

- hospital characteristics
- staffing
- resources
- geographical area
- level of specialization
- private, public

🩺 Clinical data

- health care registers
  - Mandatory (by law)
  - eg: National patient register, cancer register (diagnoses)
- quality registers
  - Optional for health care providers
  - (Mandatory within organisations joining)
  - conditions (diabetes, cancer, etc)
  - procedures (total hip arthroplasty)
  - Diagnoses, treatments, helth status, questionaires (PROM/PREM)

🤓 Individual background data

- socioeconomic data
- education
- income
- occupation
- family relations
- migration status
- Mortality data
  - date of death
  - cause of death

🗺️ Aggregated data

“Micro” vs. “macro” data.

- population data
- neighborhood characteristics
- pollution
- crime rates

Inclusion/exclusion criteria

- 👍 Defines the target study/register population

- 👎 Define exceptions to the general rules

🦿 Simple example

> “Every Swedish resident who had total hip arthroplasty performed in Sweden”

- **Include:** all ages, all hospitals, all reasons for the prothesis, all types of prosthesis
- **Exclude:** Swedish residents with surgery performed in other countries. Non-Swedish residents with the procedure performed in Sweden.

🥴 Complicated example

The National Quality Register for Ovarian Cancer

- **Inclusion**
  1. **Epithelial borderline tumours of the ovary**
     - Topography code according to ICD-O/2: C56.9.
     - Morphology code according to ICD-O/2 ≥80103 and <85900.
     - Borderline tumours with 5th digit 3 in the morphology code according to ICD-O/2 and benign behaviour flag = 3.
  2. **Epithelial ovarian cancer:**
     - Topography code according to ICD-O/2: C56.9.
     - Morphology code according to ICD-O/2 ≥80103 and <85900.
     - Malignant tumours with 5th digit 3 in the morphology code according to ICD-O/2 and benign behaviour flag blank.
  3. **Non-epithelial ovarian cancer:**
     - Topography code according to ICD-O/2: C56.9.
     - Morphology code according to ICD-O/2 ≥85903 and <95900, with the exception of mesotheliomas with ICD-O/2 codes in the interval ≥90500 and <90600.
     - Malignant tumours with digit 3 as the fifth digit in the morphology code according to ICD-O/2.
     - Exception for granulosa cell tumours, where all cases with morphology codes according to ICD-O/2 in the interval ≥86200 and ≤86223 are included.
  4. **Malignant tumours of the fallopian tube:**
     - Topography code according to ICD-O/2: C57.0.
     - Morphology code according to ICD-O/2 ≥80003 and <95900, with the exception of mesotheliomas with ICD-O/2 codes in the interval ≥90500 and <90600.
     - Malignant tumours with digit 3 as the fifth digit in the morphology code according to ICD-O/2.
- **Exclusion**
  - **Epithelial ovarian cancer and borderline tumours of the ovary**
    Cases with **behaviour codes 0, 1, 2, 6, or 9 as the fifth digit** in the ICD-O/2 morphology

  code are excluded.
  Morphology codes according to ICD-O/2 **<80103 and ≥85900** are excluded.
  - **Non-epithelial ovarian cancer**
    Cases with **digits 0, 1, 2, 6, or 9 as the fifth digit** in the ICD-O/2 morphology code are excluded, **with the exception of granulosa cell tumours**, for which cases with ICD-O/2 morphology codes in the interval ≥**86200 and** ≤**86223** are included even when the final digit is **0, 1, 2, or 3**.
    Morphology codes according to ICD-O/2 **<85903**, as well as codes in the intervals ≥**90500 and** <**90600** (mesotheliomas) and ≥**95900**, are excluded.
  - **Tumours of the fallopian tube**
    Cases with **behaviour codes 0, 1, 2, 6, or 9 as the fifth digit** in the ICD-O/2 morphology code are excluded.
    Morphology codes according to ICD-O/2 in the intervals ≥**90500 and** <**90600** (mesotheliomas) and ≥**95900** are excluded.
  - **For all diagnoses**, cases are excluded if the diagnosis is based solely on:
    - clinical examination (**basis of diagnosis 1**),
    - imaging procedures including radiography, scintigraphy, ultrasound, MRI, CT (or equivalent examinations) (**basis of diagnosis 2**),
    - autopsy with or without histopathological examination (**basis of diagnosis 4 or 7**),
    - surgery without histopathological examination (**basis of diagnosis 6**), or
    - other laboratory investigations (**basis of diagnosis 8**).
    - cases with **age** <**18 years** are excluded.

Coverage and completeness

- 🏥 **Institutional coverage**: proportion of all eligible units/clinics that are connected to the registry
  - e.g., 90% of hospitals performing the procedure are connected
  - Should be known by the "register holder"
- 🤒 **Case coverage**: proportion of patients who should have been reported from connected units that are actually included
  - e.g., 85% of eligible patients registered
  - The aim is to use 100 % but this is not always possible
- **Data completeness**: proportion of required data fields that are filled in for the registered patients
  - 🚬 e.g., 95% of patients have smoking status recorded
  - 🩸 e.g., 80% of patients have blood pressure data available

What is recorded?

- 👩‍⚖️ Some registers are mandated by law and regulations
- Quality registers often have a steering committee and register holder
- Reseasrh initiated databases according to specific protocols

Data linking

- Unique personal identifier
  - Not in every country!
  - Social security number similair purpose but not as widely used
- study specific id number
- HSA ("Hälso- och sjukvårdens adressregister" for staff and organisations)

Unique personal identifier

(Swedish: personnummer, reading: [1])

121212-1212 Tolvan Tolvansson

- 10 (or 12) digits
- date of birth-4 digits
- assigned at birth or immigration
- used in all health care contacts
- used for all administrative data
- sometimes reused after death
- sometimes changed (uncommon)
- sometimes inclusion criteria for register
- similair in the Nordic countries
  - Denmark: CPR number
  - Norway: Fødselsnummer
  - Finland: Henkilötunnus
  - Iceland: Kennitala

Combining data

- Similar registries in different areas/regions/countries
  - Different individuals but similar data
- Same definitions and variables?
- Same inclusion criteria?
- Don't get fooled by similar names!
- Differences and similarities within the Nordic countries [2]

Working with health care data

A lot to do before the statistical analysis!

- **Legalities**
  - Do I have the right to access this data?
  - What am I allowed to do?
  - What am I not allaowed to do?
- **Data management**

  - large datasets
  - multiple datasets
  - different formats
  - missing data
  - data cleaning
  - data transformation
  - data wrangling
  - data munging
  - data governance
  - data engineering
- **Planning**
  - What is the purpose?
  - How can I achieve my goals?
  - What if I change my plans later?
  - Can I redo my analysis?
  - How do I present/communicate my results?

R as a tool but …

- Large files often comes exported from SAS (initially "Statistical Analysis System")
- Comma-Separated Values (csv) or text files
- Application Programming Interface (API) calls
- Structured Query Language (SQL) databases
- Hierarchical data structures (eXtensible Markup Language, XML; JavaScript Object Notation, JSON, …)

Our use of R

- `{data.table}` to handle large data sets efficiently
- `{targets}` to streamline a reproducible pipeline
- `Git` for version control
- `GitHub` for colaboration
- `Quarto` for reporting

Bibliography

[1] J. F. Ludvigsson, P. Otterblad-Olausson, B. U. Pettersson, and A. Ekbom, "The Swedish personal identity number: Possibilities and pitfalls in healthcare and medical research," *European Journal of Epidemiology*, vol. 24, no. 11, pp. 659–667, 2009, doi: 10.1007/s10654-009-9350-y.

[2] K. Laugesen *et al.*, "Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries," *Clinical Epidemiology*, pp. 533–554, Jul. 2021, doi: 10.2147/CLEP.S314959.