Handouts

Updated handouts

- This document is generated automatically and contains all lecture slides.
- When modifications are made to any of the lecture slides, this should be reflected here automatically.
- If you are viewing the HTML version of the handouts, a static PDF might also be downloaded by clicking `Other formats > Typst` in the right side menu of the page.
- If you are reading the PDF version of the document, the HTML version is availble here

Those handouts where last modified: **2026-01-21**.

EL1: Intro

Lecture Slides

Associated litterature (References at the end)

- [1, ch. 1]
- [2]
- [3]

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: [2])

> 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 [3]

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

EL2: European legislation

Lecture Slides

> Associated litterature (References at the end)
>
> - [4]
> - [1, ch. 4]

Legal part

- Today: European legislation (mainly GDPR!)
  - Lecture handouts main sourse for examination (seminair ES1 and possibly DISA exam).
  - Article [4]. Focus on the introduction, the section "GDPR-related enablers and barriers to cross-country health data exchange in Europe" (in the results section incl. figures and tables), discussion and conclusions. Examined as part of seminair ES1 (not the DISA exam).
- Next lecture: Swedish legislation (including associated reading). Examined as part of seminair ES1 (not the DISA exam).
- Consequence: [1, ch. 4]. Read the beginning. Skip "Medical Information Mart for Intensive Care". Read the "Synthea" section. The "Synthea" section does not have a legal focus but the legal parts explains why we use this data. Read for your own understanding (not examied).

European legislation

- **GDPR** (our focus)
  - Defines the legal conditions for **processing personal data**
  - Focuses on protection, safeguards, and accountability
- **European Health Data Space (EHDS)**
  - Establishes a European framework for access to health data for research, statistics, and policy.
  - Focuses on data access, governance, and interoperability
  - Increases opportunities for cross-national health statistics
- **EU Data Act**
  - Regulates who may access data and under what conditions, across sectors.
  - Indirectly relevant for health statistics through device-generated and digital service data.

> ! Important
>
> Legal and governance frameworks enable access to data, but **statistical expertise remains essential** for ensuring data quality, valid inference, and meaningful interpretation.

EU law vs Swedish law

- EU legislation tends to be more detailed in the legal text itself

- This is because EU law must be:
  - applied uniformly across many different legal systems
  - interpreted without relying on national preparatory works
- Interpretation of EU law relies mainly on:
  - the wording of the legislation
  - recitals (non-binding explanaitions before the articles describing the purpose and context).
- Swedish legislation is often:
  - shorter and less detailed in the statutory text
  - supplemented by extensive **preparatory works (förarbeten)**
- In Sweden, preparatory works are a central interpretative source for courts and authorities

➡️ The difference reflects **different legislative techniques**, not necessarily a difference in regulatory ambition.

> ℹ Source
>
> The GDPR is available in all official EU languages via EUR-Lex. Take a quick look to get a very brief overview. However, it is recommended reading only if you suffer from insomnia — it is not required for fulfilling the course requirements!

GDPR

- Regulation (EU) 2016/679 (GDPR)
- Enforced since May 25, 2018
- Regulates the **processing** of **personal data**
- Aims to protect the privacy and rights of individuals
- Sets out rules for data **controllers** and **processors**

European Union (EU)

- GDPR applies **directly and uniformly** as law
- No national implementation required
- Member States may:
  - introduce **supplementary legislation**
  - allow legal exceptions, e.g. for:
    - research
    - public interest
    - health data

European Economic Area (EEA)

Countries: Norway, Iceland and Liechtenstein

- GDPR applies via the **EEA Agreement**
- Implemented into national law
- In practice:
  - very similar application as within the EU

- same core principles, rights, and obligations

United Kingdom (UK)

- EU GDPR no longer applies directly after Brexit
- Replaced by:
  - UK GDPR
  - Data Protection Act 2018

Switzerland

- Not part of EU or EEA
- GDPR does **not** apply as law
- Instead: Federal Act on Data Protection (FADP)
  - Revised to align closely with GDPR

International laws

- Note that other countries have different laws and regulations
- In USA, for example, HIPAA regulates the use and disclosure of protected health information (PHI)
  - Different states have different laws as well
- When collaborating internationally, compliance with all relevant laws is required

Definitions

GDPR article 4:

Personal data

> means **any information** relating to an **identified or identifiable natural person** (*data subject*); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Processing

> means **any operation** or set of operations which is **performed on personal data** or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

Pseudonymisation

> means the processing of personal data in such a manner that the personal data can **no longer be attributed to a specific data subject without the use of additional information**, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

Controller

means the natural or legal person, public authority, agency or other body which, alone or jointly with others, **determines the purposes and means** of the processing of personal data […]

Processor

means a natural or legal person, public authority, agency or other body which processes personal data **on behalf of the controller**;

Consent of the data subject

means any **freely given, specific, informed and unambiguous** indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies **agreement to the processing** of personal data relating to him or her;

Personal data breach

means a breach of security leading to the accidental or unlawful destruction, loss, alteration, **unauthorised disclosure of, or access to, personal data** transmitted, stored or otherwise processed;

Data concerning health

means personal data related to the **physical or mental health** of a natural person, including the provision of **health care services**, which reveal information about his or her health status;

Legal grounds for processing personal data:

GDPR article 6 (1):

Processing shall be lawful only if and to the extent that at least one of the following applies:

a. the data subject has given **consent** to the processing of his or her personal data for one or more specific purposes; ~~processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;~~
b. ~~processing is necessary for compliance with a legal obligation to which the controller is subject;~~
c. ~~processing is necessary in order to protect the vital interests of the data subject or of another natural person;~~
d. processing is necessary for the performance of a task carried out in the **public interest** or in the exercise of official authority vested in the controller;
e. processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.
f. ~~processing is necessary for compliance with a legal obligation to which the controller is subject under Union or Member State law requiring the processing of personal data for a specific purpose.~~

Legal ground (d) is the most relevant if you work iwith secondary data in the public sector (research and reporting etc). (a) is relevant to collect primary data for research etc. (e) is a delicate one ...

Processing of special categories of personal data

GDPR Article 9 (1):

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, **data concerning health** or data concerning a natural person's sex life or sexual orientation shall be **prohibited**. 😱

But ...

> Paragraph 1 shall not apply if one of the following applies:

- (a) the data subject has given **explicit consent** to the processing of those personal data for one or more specified purposes [...]
- (i) processing is necessary for reasons of **public interest** in the area of **public health**, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices [...]
- (j) processing is necessary for archiving purposes in the public interest, **scientific** or historical research purposes or **statistical purposes** (🥳) in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

Safeguards

GDPR article 89:

> Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or **statistical purposes**

1. Processing for archiving purposes in the public interest, scientific or historical research purposes or **statistical purposes**, shall be subject to **appropriate safeguards**, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that **technical and organisational measures are in place** in particular in order to ensure respect for the principle of **data minimisation**. Those measures may include **pseudonymisation** provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

2. Where personal data are processed for scientific or historical research purposes or **statistical purposes**, Union or Member State law may provide for derogations [...] so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

Technical Safeguards

Examples of technical safeguards include:

- Pseudonymisation
- Encryption of personal data
- Access controls and authentication
- Logging and monitoring of access
- Secure storage and transmission
- Use of secure software environments often enforced by organizational standards

Organisational Safeguards

Examples of organisational safeguards include:

- Defined roles and responsibilities
- Internal policies and procedures
- Staff training and confidentiality obligations
- Data protection by design and by default
- Incident and breach response procedures
- Documentation and accountability measures
- Working for a health care organization might require an agreement of secrecy

Data Minimisation and Purpose Limitation

- Only data that are **necessary** should be processed
- Data are processed **only for specified purposes**
- Access is limited to authorised personnel
- Retention periods are defined and respected

Pseudonymisation and Anonymisation

- **Pseudonymisation** reduces risks while allowing reuse of the data
  - Identifiers are kept separately and protected
- **Anonymisation** removes data from GDPR scope (if irreversible)

> ! Important
>
> - Pseudonymisation $\neq$ Anonymisation
> - Removing the Swedish personal identification number (PIN) is not a guarantee for pseudonymisation

Data Controller

- The entity that **determines the purposes and means** of the processing of personal data
- Bears the **primary legal responsibility**
- Responsible for:
  - Lawful basis
  - Compliance with GDPR principles
  - Transparency and information to data subjects
  - Appropriate technical and organisational measures
- Typical examples:
  - Public authorities
  - Universities
  - Regions and municipalities

Data Processor (PUB) under GDPR

- Processes personal data **on behalf of the controller**
- Acts **only on documented instructions** from the controller
- May **not** determine purposes of processing
- Has direct responsibilities for:
  - Security of processing (Article 32)
  - Confidentiality
- Must be governed by a **data processing agreement**

Controller–Processor Relationship

- A formal **Data Processing Agreement (DPA)** is required
- The agreement must specify:
  - Subject matter and duration
  - Nature and purpose of processing
  - Types of personal data
  - Categories of data subjects (patients, students, citizens, …)
  - Security measures
- The controller remains responsible even when processing is outsourced

Example: Sahlgrenska

- If a researcher work at the Sahlgrenska university hospital, VGR might be the data controller (personuppgiftsansvarig; PUA)
- If he/she asks for statistical consulting from the Sahlgrenska Academy, GU might be the data processor (personuppgiftsbiträde; PUB)

European Health Data Space (EHDS)

What is it?

- EHDS is an EU-wide legal and technical framework for the use and sharing of health data
- It aims to:
  - improve access to health data across borders

  - support healthcare, research, **statistics**, and policy-making
- Focuses on data access and governance

Two Main Pillars

- Primary use of health data
  - Use of data for individual patient care
  - Cross-border access to electronic health records
- Secondary use of health data for
  - statistics
  - scientific research
  - public health
  - policy evaluation and innovation

Implementation Timeline

- **2025**: EHDS regulation enters into force
- **2025–2027**: Development of implementing and technical acts
- **From ~2029 onwards**:
  - national infrastructures become operational
  - cross-border access for secondary use starts to function in practice

How EHDS Relates to GDPR

- EHDS does **not replace GDPR**
- GDPR continues to govern:
  - personal data protection
  - lawful bases
  - safeguards for health data
- EHDS provides **procedures and structures** for lawful data access under GDPR

➡️ GDPR defines *whether* data may be processed

➡️ EHDS defines *how* data can be made available

Why EHDS Matters for Statisticians

- EHDS explicitly recognises **statistics** as a legitimate purpose
- It facilitates access to:
  - large-scale health datasets
  - cross-national data sources
- It increases demand for:
  - data quality assessment
  - metadata interpretation
  - harmonisation and comparability analyses

EHDS Does *Not* Do This

- EHDS does not:
  - define statistical methods

  - ensure data quality automatically
  - guarantee comparability across countries
- Legal and technical access ≠ valid statistical inference

The EU Data Act

What Is It?

- The Data Act is an EU regulation on access to and sharing of data
- Focuses mainly on:
  - data generated by connected products and digital services (IoT)
  - business-to-business (B2B) and business-to-government (B2G) data sharing
- It is **not a data protection regulation**

➡️ The Data Act is about *who may access data and under what conditions.*

How the Data Act Relates to Health Data

- The Data Act does not primarily target health registers
- However, it may affect:
  - data generated by medical devices
  - digital health services
  - health-related IoT data

➡️ Health data may fall under the Data Act depending on how it is generated.

EL2: Swedish legislation

Lecture Slides

> Associated litterature (References at the end)
>
> - [5]

Swedish law

> Swedish law
>
> The Swedish legislation will be discussed as an example. It will be examined through discussions in a later seminar (where you will have access to the information below). You are not required to memorize individual laws or specific legal provisions.

Backgrund of Swedish law

Optional reading

- **Fundamental laws** (“grundlagar”) decided by the parliament but stable over time

  - Not a “constitution”
- **Ordinary laws** (parliament)
  - New ones all the time
  - New laws to update existing laws
- Published in Svensk författningsamling, SFS
  - The “big blue book” is only a smaller collection of important laws
- **ordinances** (“förordning”) from the goverment to implement laws
- **regulations** (“föreskrifter”) from authorities to implement laws and ordinances

Two main branches of law

- Civil and criminal law:
  - state what you can not do (everything else is “legal”)
  - Handled by ordinary courts
  - Ex: Brottsbalken kap 20 om tjänstefel m.m. (The Swedish Penal Code (Brottsbalken), Chapter 20 — Offences Relating to Public Office.)
- Public law: relationship between individuals and the state etc
  - state what the public authorities must and can do (everything else is “illegal”)
  - Handled by administrative courts
  - What we mostly care of here

Tryckfrihetsförordningen (TF)

- World’s oldest freedom of the press law (since 1766)
- Chapter 2: Public access to official documents
- Applies to public authorities and institutions
  - Including health care registers and medical records held by public authorities
- Everyone has the right to access official documents (TF 2.1)
- but there are exceptions (TF 2.2)
  - e.g., if disclosure would violate privacy or national security

- if so, the goverment has the right to provide ordinary laws that restrict access (which they do ...)

Offentlighets- och sekretesslagen (OSL)

- Law that regulates public access to official documents and secrecy
- Applies to public authorities and institutions
- Defines what information is considered secret and under what circumstances

OSL Chap 21

Secrecy for private individuals' personal circumstances **no mather the context**

- E.g., health data, economic circumstances, family relations

> OFS 21.1: Secrecy applies to information concerning an individual's health or sexual life, such as information about illnesses, substance abuse, sexual orientation, gender reassignment, sexual offences, or other similar information, **if it can be assumed** that disclosure of the information would **cause significant harm** to the individual or to someone closely related to them.

OSL Chap 24

Secrecy for the protection of individuals in research and statistics.

- Special research databases etc
- Some regulations for research ethics boards

OSL Chapter 25

Secrecy for the protection of individuals in activities relating to **health and medical care** etc.

> OFS 25.1: Within the health and medical care services, secrecy applies to information concerning an individual's state of health or other personal circumstances, **unless it is clear** that the information may be disclosed **without causing harm** to the individual or to someone closely related to them. The same applies to other medical activities, such as forensic medical and forensic psychiatric examinations, insemination, in vitro fertilisation, abortion, sterilisation, circumcision, and measures to prevent communicable diseases.

- Exceptions exists,
  - for example to submit medical patient data to quality registers
  - to share data between public organisations for research purposes or statistics (OFS 25.11 p. 5).

OSL Chapter 10

Provisions on disclosure overriding secrecy and provisions on exemptions from secrecy

> OFS 10.28: Secrecy does not prevent information from being disclosed to another authority where a duty to provide information follows from an act or an ordinance.

- This would apply to data sharing for research purposes when there is a legal basis for that

Patientdatalagen (PDL)

The Patient Data Act (PDL)

- regulates the processing of personal data within **health and medical care** in Sweden.
- Applies to **healthcare providers** (public and private).
- Main objectives:
  - Protect patient privacy
  - Ensure safe and effective healthcare
  - Enable **secondary use** of health data under strict conditions

Chapter 7 PDL

National and regionel quality registers

Opt-out for patients (every one is included by default until they opt out)

> PDL 7.4: Personal data in national and regional quality registers may be processed for the purpose of systematically and continuously developing and ensuring the quality of healthcare.

PDL 7.5: Personal data processed for the purposes set out in Section 4 may also be processed for the purposes of

- the production of **statistics**,
- estimating numbers for the planning of clinical research,
- **research within health and medical care**,
- disclosure to a party that will use the data for purposes referred to in Sections 1 and 3 or in Section 4, and
- …

Lag om hälsodataregister (SFS 1998:543)

This law regulates health data registers outside the health and medical care system. A new law is being proposed to replace this one.

§ 3: Personal data in a health data register may be processed for for the following purposes:

- the production of statistics,
- follow-up, evaluation and quality assurance of health and medical care, and
- research and epidemiological studies

Specific registers

Register (Swedish)	Register (English)	Governing act / ordinance
Folkbokföringen	Population Register	Population Registration Act (1991:481); Population Registration Ordinance (1991:749)
Totalbefolkningsregistret (RTB)	Total Population Register	Official Statistics Act (2001:99); Official Statistics Ordinance (2001:100)
Nationella patientregistret	National Patient Register	Health Data Act (1998:543); Ordinance on the National Patient Register (2001:707)
Cancerregistret	Swedish Cancer Register	Health Data Act (1998:543); Cancer Register Ordinance (2001:709)
Dödsorsaksregistret	Cause of Death Register	Health Data Act (1998:543); Cause of Death Register Ordinance (2001:709)
Läkemedelsregistret	Prescribed Drug Register	Act on the Prescribed Drug Register (2005:258); Ordinance (2005:363)
Medicinska födelseregistret	Medical Birth Register	Health Data Act (1998:543); Medical Birth Register Ordinance (2001:708)
Tandhälsoregistret	Dental Health Register	Health Data Act (1998:543); Dental Health Register Ordinance (2008:194)

Arkivdatalagen (ADL)

- Regulates the management of public records and archives
- Applies to public authorities and institutions
- Differnt authorities then have different rules for how long data must be kept
  - For example healthcare data is often required to be kept for at least 10 years

GDPR and Swedish law

- GDPR is directly applicable in Sweden
- There are refereces to GDPR in Swedish laws such as PDL and OSL
- Swedish laws may provide additional regulations and requirements beyond GDPR
- Data protection authorities in Sweden: Integritetsskyddsmyndigheten (IMY)
- Should be easy to collaborate across EU borders due to GDPR, but more difficult with non-EU countries

Etikprövningslagen (EPL)

- Regulates ethical review of research involving humans (including their data!)
  - Had received some critisism and might be revised
- Applies to research projects conducted in Sweden

- Requires ethical review and approval by an ethics review board
- Aims to protect the rights, safety, and well-being of research participants
- Based on the Declaration of Helsinki and other international ethical guidelines
- One application for each new research project
  - Ammendments for changes in already approved projects
- Application fees applies
- Swedish Ethical Review Authority

EL4: Tooling

Lecture Slides

> Associated litterature (References at the end)
>
> -

History

> It is widely acknowledged that the most fundamental developments in statistics in the past 60 years are driven by information technology (IT). We should not underestimate the importance of pen and paper as a form of IT but it is since people start using computers to do statistical analysis that we really changed the role statistics plays in our research as well as normal life.
>
> Although: “Let’s not kid ourselves: the most widely used piece of software for statistics is Excel.”” /Brian Ripley (2002)

Short overview

General-Purpose Programming Languages

Early statistical computing relied heavily on:

- **FORTRAN**
  - Dominant language for numerical and statistical computation
  - Statistical methods implemented as libraries and subroutines
  - Still used as subroutines in modern statistical software!
- **ALGOL / ALGOL 60**
  - Used mainly in academic environments
- **PL/I**
  - Used in some government and industrial contexts

📌 These languages required substantial programming expertise.

Early Statistical Packages

Several dedicated statistical systems emerged:

- **SPSS** (1968)
  - Originally batch-oriented
  - Widely used in social sciences
  - Originaly “Statistical Package for the Social Sciences”
- **BMDP** (1960s)
  - Developed at UCLA
  - Common in medical statistics
  - Bio-Medical Data Package
- **GENSTAT** (1968)
  - Focused on agricultural statistics
- **MINITAB** (1972)
  - Designed for teaching and education
  - Still popular in quality control

SAS: A Transitional System

- **SAS** (early 1970s)
- Developed for agricultural and biostatistical analysis
- Script-based, but largely batch-oriented
- Became a standard in:
  - Government agencies
  - Large organisations
- Known for strong data management capabilities
- Still widely used in pharmaceutical industry and clinical trials

📌 SAS predates S but influenced later statistical workflows.

Limitations of Pre-S Systems

Common limitations included:

- Batch processing rather than interactivity
- Separation of data management and analysis
- Limited graphics capabilities
- High barriers to exploratory data analysis

S

- S takes form at Bell Laboratories (interactive statistical computing).
- John Chambers leads the effort.
- **1976:** first working version of S runs on GCOS
- **1979:** S2 is ported to UNIX; UNIX becomes the primary platform
- **1980:** S is first distributed outside Bell Labs
- **1981:** source versions are made available
- **1984:** key S books published (often called the “Brown Book” era)

The New S Language

- **1988:** "New S" is released (major language redesign)
- **1988: S-PLUS** is first released as a commercial implementation of S
- **1991:** *Statistical Models in S* ("White Book") popularizes formula notation (the ~ operator), data frames, and modeling workflows

R

- **1993:** first versions of **R** are published (Auckland; Ross Ihaka & Robert Gentleman)
- **1995:** R becomes open source (GPL)
- **1997:** the R Core group forms; **CRAN** is founded (Kurt Hornik)
- **2000: R 1.0.0** is released 2000-02-29

RStudio brings an IDE to the R community

- **2009:** RStudio (the company) is founded
- **2011:** RStudio IDE is introduced as an open-source IDE for R (desktop + server)

Microsoft (Revolution Analytics)

- **Jan 2015:** Microsoft announces it will acquire **Revolution Analytics**
- Microsoft promotes enterprise R offerings (e.g., Microsoft R Open / R Server)
- **2016:** SQL Server 2016 introduces **R Services** (in-database R)
- **2017:** Microsoft expands the stack under "Machine Learning Server" branding
- **June 2021:** Microsoft announces retirement of **Microsoft Machine Learning Server**
- **July 1, 2023:** Microsoft era over

RStudio becomes Posit

- **July 27, 2022:** RStudio rebrands as **Posit**
- **July 28, 2022: Quarto** is announced as a next-generation scientific and technical publishing system (multi-language, multi-engine)

Modern IDE

Positron

- New generation IDE for data science
- From Posit PBC
- Free for individual use
- Based on Code OSS (open source version of VS Code from Microsoft)
- For both R and Python (Julia?)

Quick tour

Video: https://www.youtube.com/watch?v=4Ir_HX4riHw

! Positron assistant (mentioned in the video)

This feature will most likely be disabled in any secure working environment. Such environments often have strict rules about data privacy and security, which may conflict with the assistant's functionality. Health data in SENSITIVE and SECURE environments must not be shared with external services, including AI assistants, to comply with data protection regulations and institutional policies.

It is recommended to not rely on such tools during the course (even if all our data is synthetic). If you start to rely on such tools, you might get difficulties the day you work with real data (might lead to prosecution for "brott mot tystnadsplikten" which is not only public, but actually civil law ("Brottsbalken") with prision sentence as a possibility). Society put an extreme emphasis on protecting health data, and rightfully so!

Version control

Before Version Control Systems

1950s–1970s: Early software development relied on:

- Manual file naming:
  - `analysis_final.f`
  - `analysis_final_v2.f`
- Physical media:
  - Punch cards
  - Magnetic tape
- Centralised mainframes

📌 No automated tracking of changes.

Floppy discs, Mail, and Shared Directories

1970s–1980s: Common practices included:

- Copying files to:
  - Floppy disks
  - Magnetic tapes
- Sending media by **postal mail**
- Sharing files via:
  - Network drives
  - FTP servers

📌 Version control was social, not technical.

Early Version Control Systems

1980s: First-generation tools focused on single files:

- **SCCS** (Source Code Control System, 1972)
- **RCS** (Revision Control System, 1982)

Characteristics:

- Versioning per file
- Linear history
- Central storage

Centralised Version Control

1990s: Project-level systems emerge:

- **CVS** (Concurrent Versions System)
- **Subversion (SVN)**

Key features:

- Central repository
- Multiple users
- Check-in / check-out model

📌 Still required constant access to the central server.

Limitations of Centralised Systems

Common problems:

- Single point of failure
- Poor support for branching and merging
- Difficult offline work
- Slow operations on large repositories

These limitations became critical for large projects.

Git

- **Git** is created by Linus Torvalds in 2005
- Original motivation:
  - Support Linux kernel development
  - Replace proprietary tools

Design principles:

- Distributed architecture
- Fast local operations
- Strong support for branching and merging

Distributed Version Control with Git

Key ideas in Git:

- Every clone is a full repository
- Local commits without network access

- Cheap and fast branches
- Cryptographic integrity (hash-based)

📌 Collaboration becomes more flexible and robust.

Hosting Platforms

Platforms built around Git:

- **GitHub** (2008)
- **Bitbucket** (2008)
- **GitLab** (2011)
- Gitea - open source alternative to GitHub (get the same functionallity locally or on a server)

They add:

- Pull requests / merge requests
- Issue tracking
- Code review
- CI/CD integration

Version Control Today

Modern usage includes:

- Code
- Documentation
- Data analysis (scripts, notebooks)
- Configuration and infrastructure

Git is integrated into:

- IDEs (VS Code, Positron)
- CI/CD pipelines
- Cloud platforms

Version Control Beyond Code

Today, version control supports:

- Reproducible research
- Collaborative writing
- Data science workflows
- Teaching and learning

📌 Version control is now a core professional skill.

WARNING!

- Not everything should be shared!
- Scripts and documentation yes!
- But **Health data is sensitive!**

    - Do not share it!
    - Avoid unintentional sharing!
    - Private repositories are still shared with hosting provider!
- Avoid explicit file paths and sensitive info in scripts!
- Can give information about data location and internal structure!
- No hard-coded passwords or API keys!

Git basics

(After installing the Git software)

- Collect all files related to a project in a folder
- Initialize a git repository in that folder
- Make changes to files
- Stage changes for commit
- Commit changes with a message
- Possibly push commits to remote repository

```
cd path/to/your/project
git init
git status
## make changes to files
git add filename1 filename2
git commit -m "Descriptive message about changes"
git remote add origin
```

Video tutorials

The video below is a good start to understand the basic concepts of Git and GitHub (and there are others to be found on YouTube). Note that there are a lot of videos on Git combined with Rstudio. They might still be relevant even though we are using Positron here.

Video: https://www.youtube.com/watch?v=mJ-qvsxPHpY

Inspirational video

Watch this video even though some parts might be overwhelming. It gives a good overview of the current state (2025), even though many things will be too advanced for this course (it is not specifically aimed for statisticians or R users).

Video: https://www.youtube.com/watch?v=vA5TTz6BXhY

! .gitignore

The `.gitignore` file is very important in settings with health data! Pay close attention to this section of the video!

Git in Positron

- Remember that Positron is build on Code OSS (which shares a lot of features with VS Code).
- Branching and merging is possible but we will not cover that here.

Short official introduction from Microsoft:

Video: https://www.youtube.com/watch?v=i_23KUAEtUM

More detailed introductions. Watch both! The first one is based on a Windows version of VS code and the second on Mac but the concepts are the same:

Video: https://www.youtube.com/watch?v=z5jZ9lrSpqk

Video: https://www.youtube.com/watch?v=twsYxYaQikI

> **i** Overwhelmed?
>
> This video includes some parts which might be overwhelming if you are new to Git and GitHub. Don't worry! You don't need to understand everything right away. Just try to follow along with the basic concepts and steps. You will get more comfortable with practice.

Statistics projects

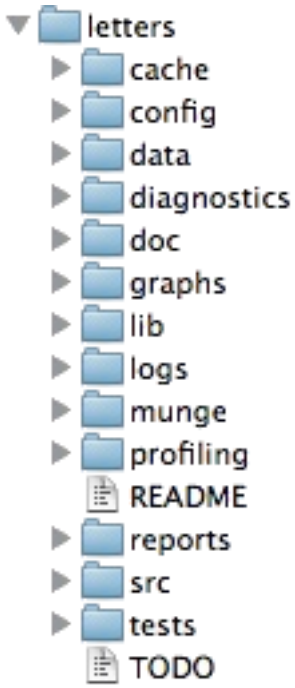
Not a single R script

- Real projects are more complex than a single R script!
- Multiple scripts
- Data files
- Documentation
- Reports
- Version control
- Reproducible workflows

Project structure

Common file structures

- Help you organize your thoughts
- Help others to collaborate
- simplifies paths used in your code



Example structure

```
/.../my_project/
  ├── README.md        - project documentation
  ├── TODO             - what should be done next?
  ├── .git             - handled by git (hidden folder)
  ├── .gitignore       - used by git but your responsibility!
  ├── data/            - your data files (not under version control!)
      ├── cancer.csv
      └── patients.qs
  ├── R/               - your saved R functions
      ├── function1.R
      └── function2.R
  ├── reports/
  └── _targets.R       - targets pipeline script
```

`README.md`

- Document the purpose of the project
- What is it about?
- What is the aim?
- Who to contact for questions?
- In what circumstances was it created?

Markdown format (simple text with some possible formatting)

`data` folder

- Store your data files here as they are when you get them
- Avoid any modifications to the raw files!
- It is very easy to forget what you do if it can not be traced by code
- Do NOT include this folder in version control!
- Git is not good at handling large files
- Sensitive data should not be shared!
- Add `data/*` to your `.gitignore` file
- In realistic projects, data might come in varying formats
  - csv, txt, xlsx, sas7bdat, sav, dta, etc
  - some files might be very big (gigabytes not uncommon)

`R` folder

- Store your R functions here
- Document their purpose inline!
- Helps you to reuse code
- Easier to read main scripts
- Easier to test and debug code
- Easier to share code between projects

`reports` folder

- Store your reports here
- Quarto documents
- Documemnt your analysis
- Include figures and tables
- Share with external collaborators

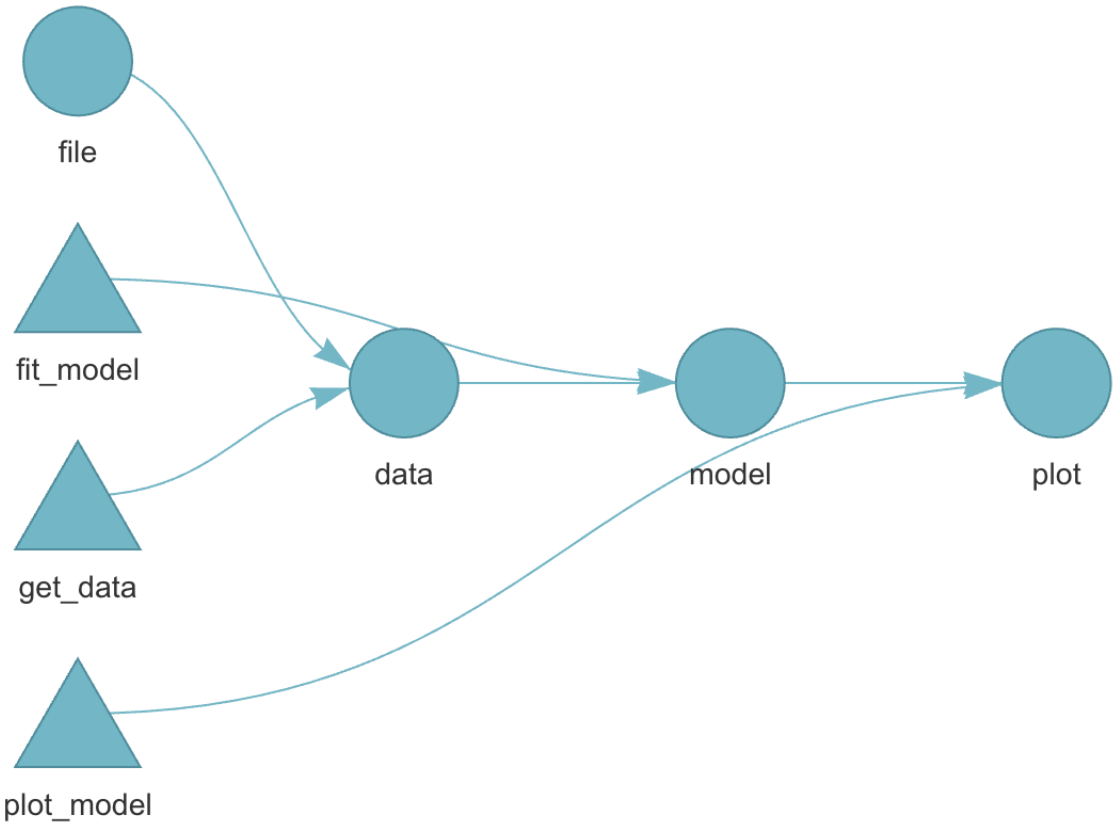
`_targets.R` file

Use the `targets` package to create a reproducible data analysis pipeline

Video: https://player.vimeo.com/video/700982360?h=38c890bd4f”

Pipeline

- Define steps in your analysis as targets
- Define dependencies between targets
- Automatically track changes and rerun only necessary parts



Why?

- You do not want to rerun everything all the time!
- You want to keep track of what you have done
- You want to be able to reproduce your results later
- You want to share your workflow with others

Reading

- A bit old but still relevant: https://happygitwithr.com/
- Targets overview
- Target manual

EL5: Analysis pipelines

Lecture Slides

Associated litterature (References at the end)

-

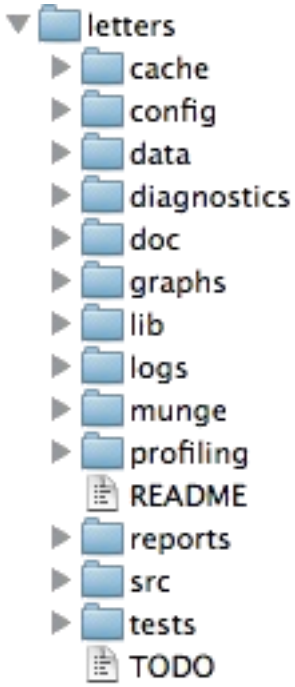
Not a single R script

- Real projects are more complex than a single R script!
- Multiple scripts
- Data files
- Documentation
- Reports
- Version control
- Reproducible workflows

Project structure

Common file structures

- Help you organize your thoughts
- Help others to collaborate
- simplifies paths used in your code



Example structure

```
/.../my_project/
  ├── README.md        - project documentation
  ├── TODO             - what should be done next?
  ├── .git             - handled by git (hidden folder)
  ├── .gitignore       - used by git but your responsibility!
  ├── data/            - your data files (not under version control!)
      ├── cancer.csv
      └── patients.qs
```

```
├── R/               - your saved R functions
    ├── function1.R
    └── function2.R
├── ...
├── ...
└── _targets.R       - targets pipeline script (WHAT??? :-))
```

`README.md`

- Document the purpose of the project
- What is it about?
- What is the aim?
- Who to contact for questions?
- In what circumstances was it created?

Markdown format (simple text with some possible formatting)

Markdown

TODO: intro and give examples + link etc

`data` folder

- Store your data files here as they are when you get them
- Avoid any modifications to the raw files!
- It is very easy to forget what you do if it can not be traced by code
- Do NOT include this folder in version control!
- Git is not good at handling large files
- Sensitive data should not be shared!
- Add `data/*` to your `.gitignore` file
- In realistic projects, data might come in varying formats
  - csv, txt, xlsx, sas7bdat, sav, dta, etc (we will cover some of these later)
  - some files might be very big (gigabytes not uncommon)

`.gitignore` fiile

TODO

`R` folder

- Store your R functions here
- You have learned about R functions in the earlier R course
- We will cover some more of that later
- Document their purpose inline!
- Helps you to reuse code
- Easier to read main scripts if functions are defined elsewhere
- Easier to test and debug code
- we eill not cover testing but some about debugging later
- Easier to share code between projects

`reports`

- Separate reports from analysis/data management but integrate them with the analysis pipeline
- Quarto documents
- Documemnt your analysis and thoughts/decisions along the way
- Include figures and tables
- To share with external collaborators

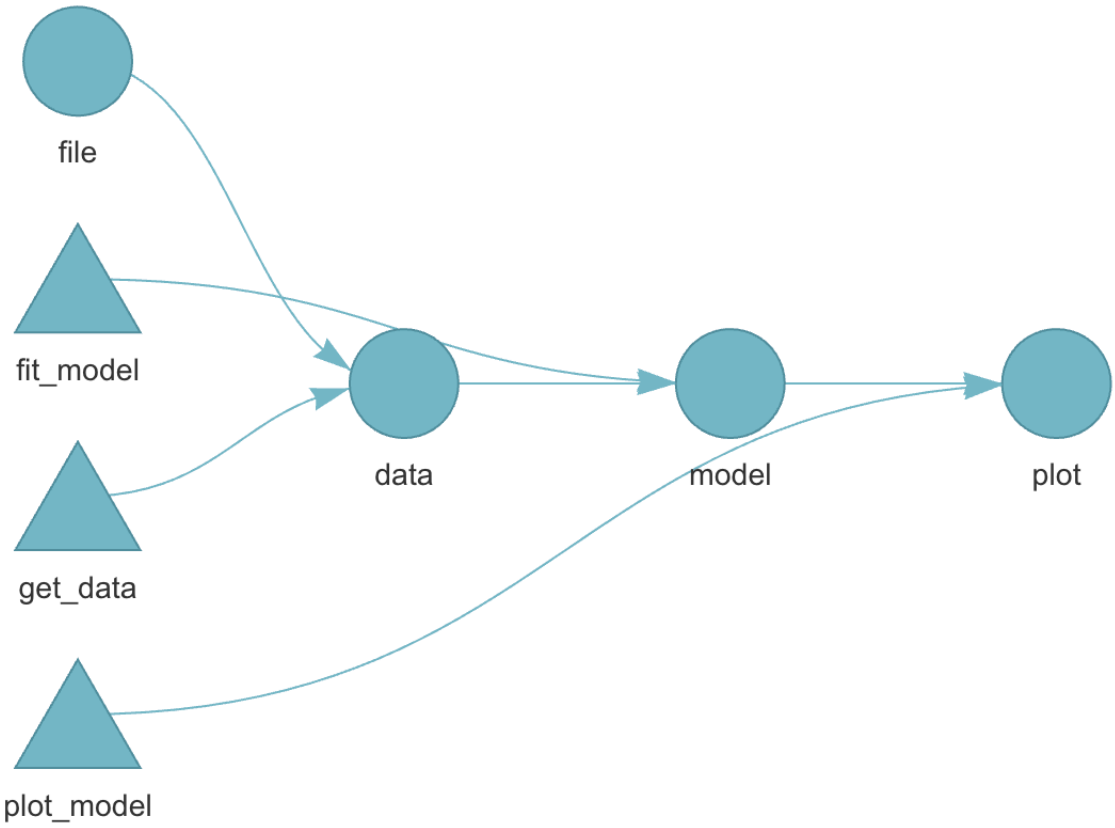
`_targets.R` file

Use the `targets` package to create a reproducible data analysis pipeline

Video: https://player.vimeo.com/video/700982360?h=38c890bd4f"

Pipeline

- Define steps in your analysis as targets
- Define dependencies between targets
- Automatically track changes and rerun only necessary parts



Why?

- You do not want to rerun everything all the time!

- You want to keep track of what you have done
- You want to be able to reproduce your results later
- You want to share your workflow with others

Reading

- Targets overview
- Target manual

EL11: Repetition

Lecture Slides

Associated litterature (References at the end)

-

NA

Complete Reading list

Articles might be obtained through the GU library (UB) or as PDF files uploaded to Canvas .

Bibliography

[1] A. Nguyen, *Hands-on healthcare data: taming the complexity of real-world data*, First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2022.

[2] 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.

[3] 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.

[4] J. Vukovic, D. Ivankovic, C. Habl, and J. Dimnjakovic, "Enablers and barriers to the secondary use of health data in Europe: general data protection regulation perspective," *Archives of Public Health*, vol. 80, no. 1, p. 115, Apr. 2022, doi: 10.1186/s13690-022-00866-7.

[5] "Public access and secrecy | Swedish National Data Service." [Online]. Available: https://snd.se/en/research-data-support/introduction-legal-aspects-research/public-access-and-secrecy