SUMMARY

Big Data for Health and Care

Summer School - 1st edition

The Journey of Data: from Collection to Impact

MAY 2023 / RESEARCH GROUP BIOMEDICAL DATA SCIENCES

UHASSELT KNOWLEDGE IN ACTION DOCTORAL SCHOOLS

Vlaamse overheid

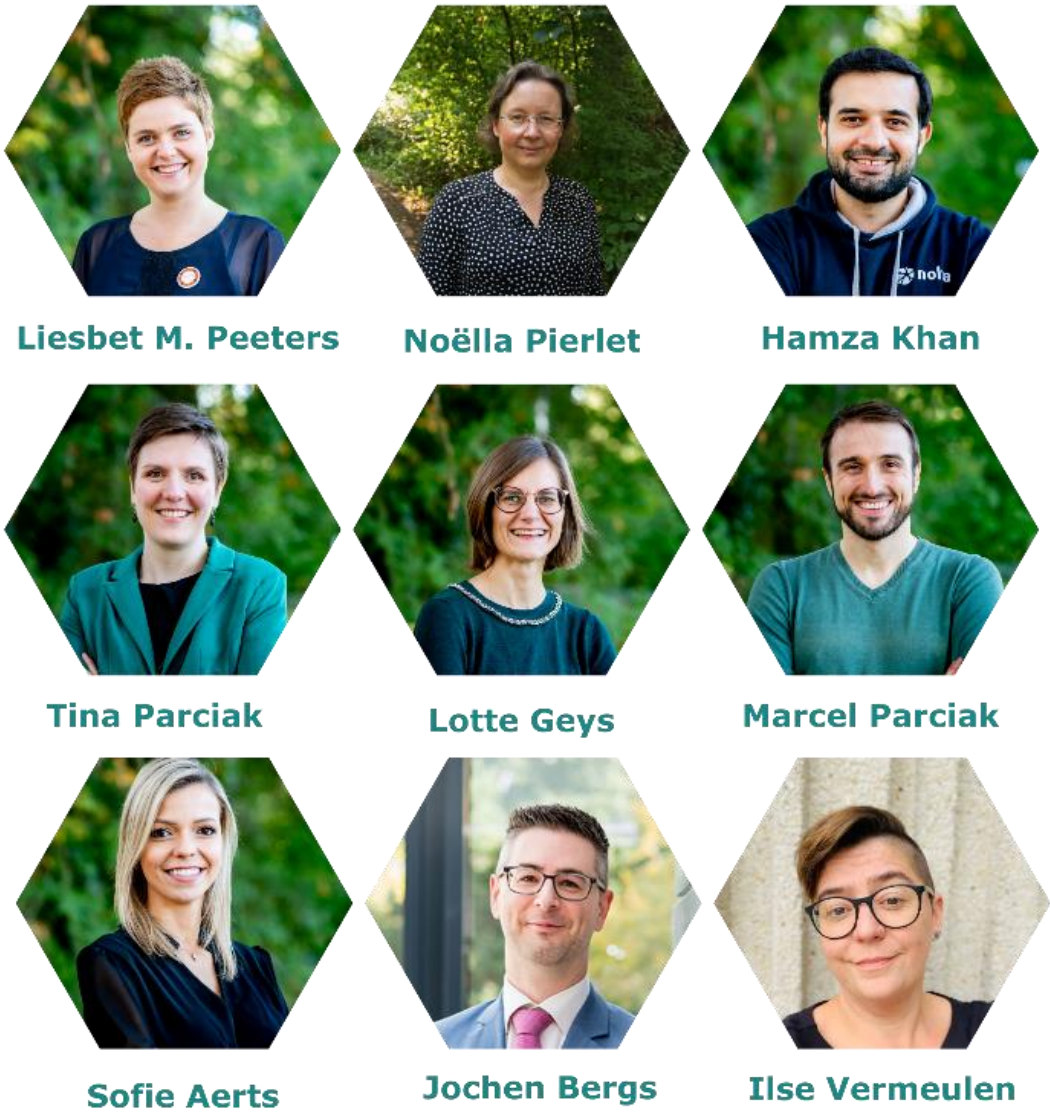
VAIA

Table Of Contents

Table Of Contents

The use of big data has the potential to revolutionize the way we approach healthcare. With the increasing availability of healthcare data, we can now leverage this data to gain new insights into disease prevention, diagnosis, and treatment. Our Summer School on "Big Data in Health and Care" provides an excellent opportunity to gain a comprehensive understanding of these topics, from data acquisition and preprocessing to analysis and governance.

Our Speakers



Visit our website to learn more about our Research Group.

SESSION I: DATA SAVES LIVES

The Promise of using Big Data in Health & Care

Session I: Data Saves Lives

The promise of using Big Data in Health & Care

We kick-started the summer school by exploring the transformative potential of big data in healthcare. Big data was introduced by some of its properties: velocity, volume, variety and veracity. We hope you have found your own inspiring #DataSavesLives story (=Value).

Data velocity

Data velocity refers to the speed in which data is generated, distributed and collected.

- **Data processing** refers to the series of actions taken to transform raw data into useful information that can be used for decision-making, analysis, or other purposes. These actions may include data collection, validation, cleaning, analysis, and storage. Data processing can be done manually or with the help of computers and software programs.

- **Data processing speed** refers to the rate at which data can be processed by a computer or a system. It is usually measured in terms of data throughput, which is the amount of data that can be processed per unit of time. Faster data processing speed means that data can be analyzed and transformed into useful information more quickly, allowing for faster decision-making and more efficient operations. Factors that can affect data processing speed include the processing power of the computer or system, the efficiency of the software used, and the complexity of the data being processed.

Data Variety

Variety is defined as the diversity of data in a data collection of problem space. Diversity can be seen as 'different sources of data', 'different formats',

Data format refers to the structure or layout in which data is stored, organized, and presented. It defines how data is arranged, encoded, and represented, and it determines how it can be interpreted and used by different systems and applications. Data can be stored and presented in various formats, such as text, numbers, images, audio, video, or a combination of these. The format used depends on the type of data, the application or system that will use it, and the intended use of the data. Common data formats include CSV (Comma Separated Values), JSON (JavaScript Object Notation), XML (Extensible Markup Language), HTML (Hypertext Markup Language), and PDF (Portable Document Format). Each of these formats has its own advantages and limitations, and the choice of format depends on the requirements of the application or system that will use the data.

Data Volume

In the context of data, volume refers to the amount or size of data that is being generated, collected, processed, stored or analyzed. It can be measured in different units such as bytes, megabytes, gigabytes, terabytes, petabytes, exabytes and zettabytes. The volume of data can range from small amounts of data such as a few kilobytes to massive amounts of data such as several petabytes or even exabytes. With the advent of big data, the volume of data that organizations are handling is increasing rapidly, and this is one of the reasons why data management and analysis have become more challenging.

There are many different types of data sources in healthcare, and the volume of data they generate can vary greatly. Here are some examples:

1. **Electronic Health Records (EHRs):** EHRs contain a patient's medical history, diagnoses, treatments, and other health-related information. The volume of data in an EHR can vary depending on the patient's medical history and the number of encounters they have had with healthcare providers. Some EHRs can contain gigabytes of data for a single patient.

2. **Medical imaging:** Medical imaging includes X-rays, CT scans, MRIs, and other types of diagnostic images. The volume of data generated by medical images can be quite large. The resulting file size varies considerably between each of these modalities. CT, MR, and the US scanned files are on average in the range of 100 to 600 kilobytes (KB), while images of modalities such as Mammography (MG) and CR can reach file sizes in the range of 27 to 30 megabytes (MB) per image

Data Veracity

Veracity means that the data is accurate, precise and trustworthy. It is important to ensure that the data used for analysis is correct and unbiased to avoid making incorrect conclusions or decisions. Veracity involves assessing the quality of the data and determining whether it can be trusted to be used for analysis.

There are several factors that can affect the veracity of data. Examples include:

1. Data quality: The quality of the data is a critical factor that affects its veracity. The data must be accurate, complete, and consistent to be trusted.
2. Data source: The source of the data can also affect its veracity. Data from reliable sources is more likely to be trustworthy than data from unreliable sources.
3. Data processing: The way the data is processed can also affect its veracity. Errors or biases introduced during data processing can lead to inaccurate conclusions.

Some examples of problems that can occur when handling data of low veracity/data quality:

- **Biases**: an organization makes a decision using a calculated value that suffers from statistical bias
- **Data lineage** (=metadata that explains where data came from and how it was calculated): an organization gets data from hundreds of sources. It discovers that one of the sources is extremely inaccurate and lacks the data lineage information to identify where the data has been stored in various databases
- **Bugs**: a software bug causes data to be calculated or transformed incorrectly
- **Noise**: what is a clinically meaningful signal versus what is just noise
- **Information security**: an organization's data is changed by an advanced persistent threat
- …

Interesting links to learn more

- simplicable.com
- xenonstack.com

SESSION II: BREAKING BARRIERS

Exploring the challenges related to Big Data in Health and Care






Session II: Breaking barriers

Exploring the challenges related to Big Data in Health and Care

Real-world health data refers to health-related information collected from various sources outside of controlled clinical trials or experimental settings. This data can come from a variety of sources such as electronic health records, claims data, patient-generated data, social media, mobile health devices, and wearable sensors. Real-world health data (RWD) provides a more comprehensive understanding of health and disease in the general population, as opposed to data collected from a limited sample of individuals in a clinical trial. It can be used to inform healthcare decisions, identify health trends and disparities, evaluate the effectiveness of treatments, and develop public health policies. Examples of RWD include patient demographics, medical history, laboratory test results, medication use, healthcare utilization, and health outcomes.

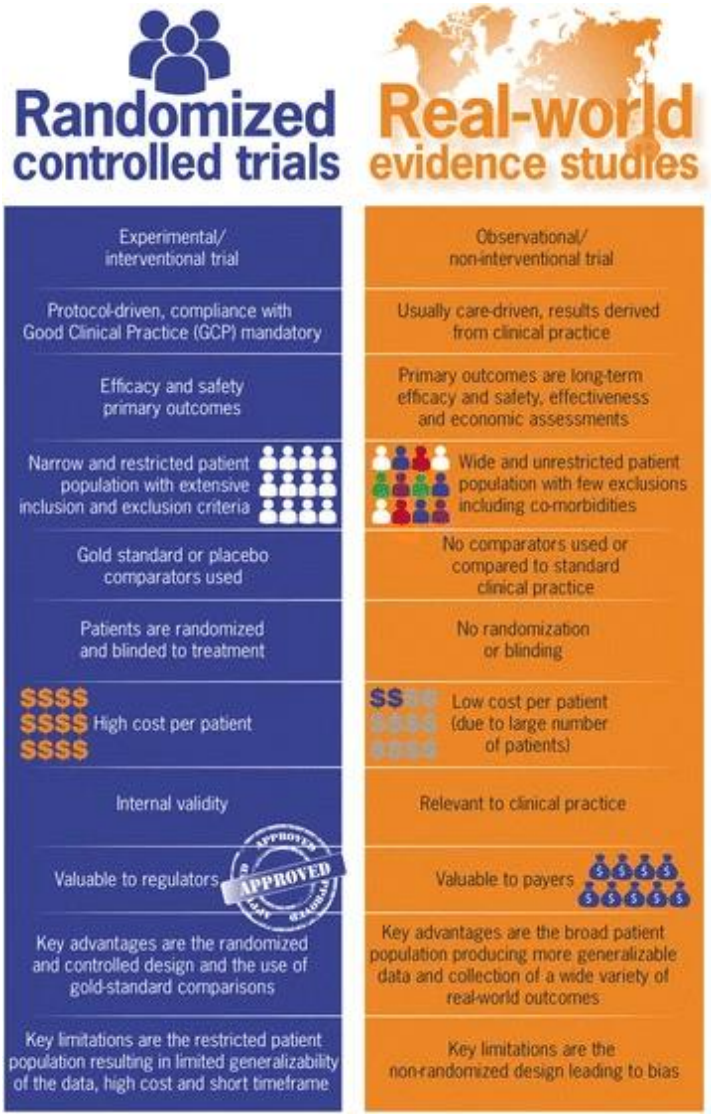
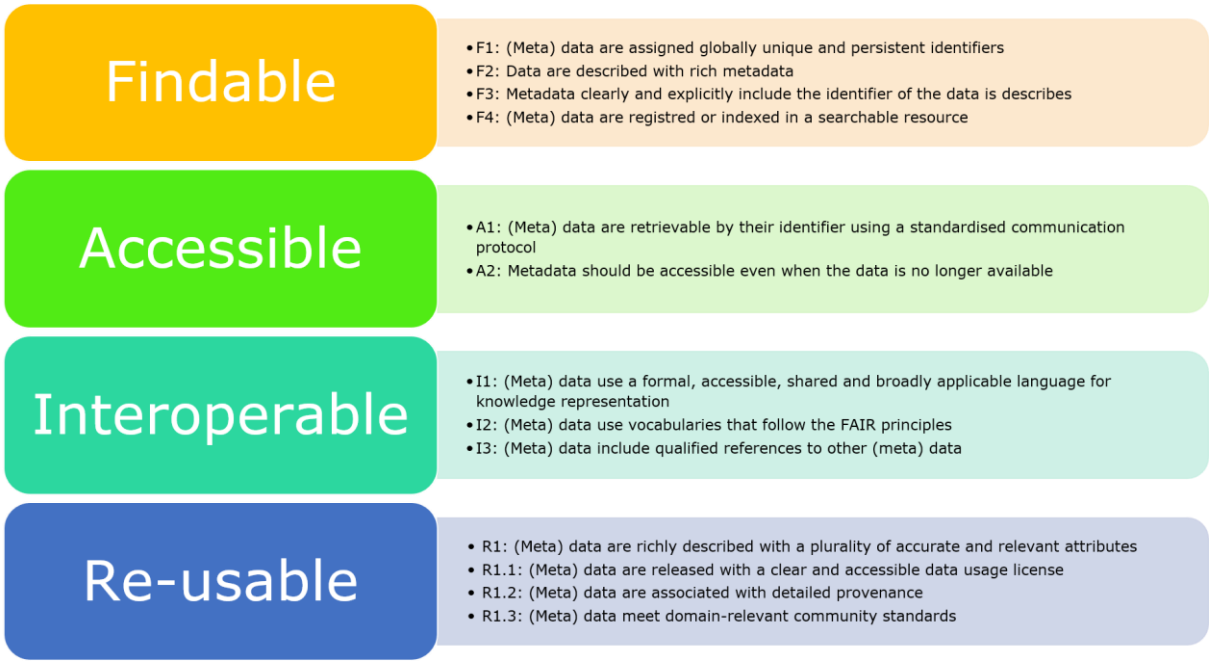


Figure: highlighted differences between randomized clinical trials and real-world evidence studies (ref: Ziemssen et al. 2016 The importance of collecting structured clinical information on multiple sclerosis. BMC Medicine)

We introduced the concept of "FAIR"



Wilkinson et al. 2016. Scientific Data

Finally - we reflected on the arisal of so-called 'data spaces'.

Key characteristics of a data space are:

- A secure and privacy preserving IT infrastructure to pool, access, process, use and share data.
- A data governance mechanism, comprising a set of rules of legislative, administrative and contractual nature that determine the rights to access, process, use and share data in a trustful and transparent manner.
- Data holders are in control of who can have access to their data, for which purpose and under which conditions it can be used.
- Presence of vast amounts of data that can be reused under certain conditions against remuneration or for free, depending on the data holder's decision.
- Participation by an open number of organizations / individuals.

In order to unleash the full potential of health data, the European Commission is presenting a regulation to set up the "European Health Data Space". This proposal:

- Supports individuals to take control of their own health data
- Supports the use of health data for better healthcare delivery, better research, innovation and policy making and
- Enables the EU to make full use of the potential offered by a safe and secure exchange, use and reuse of health data

SESSION III: MASTERING DATA INTEGRATION

How to make data fit-for-use?

Session III: Mastering data integration

How to make data fit-for-use?

The data integration process is mostly described as ETL. This stands for '**extract-transform-load**' and is a process used to integrate data from various sources into a target system, such as a data warehouse or a data lake. ETL involves extracting data from source systems, cleaning and transforming it into a suitable format, and loading it into a destination system. This process is crucial for organizations to ensure they have accurate, reliable, and timely data for their business needs.

The extraction phase involves retrieving data from source systems, which can be databases, flat files, APIs, or web services. The extraction process should be designed to retrieve only the relevant data needed for the target system, and it should be done in a way that does not impact the performance of the source systems.

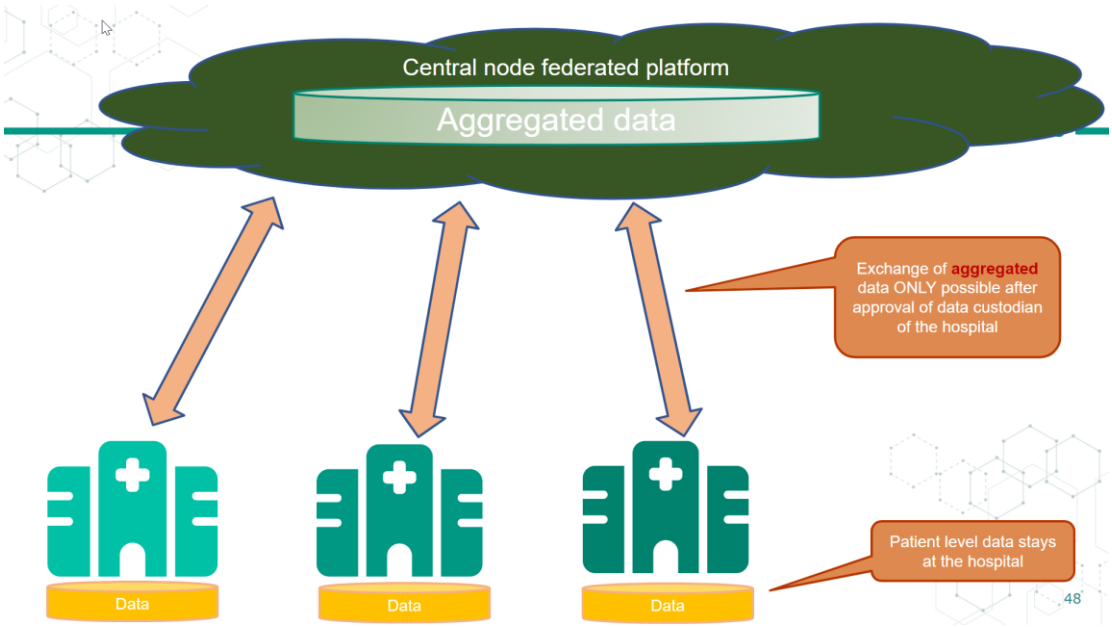
The cleaning phase is the process of identifying and correcting or removing errors, inconsistencies, and duplicates from the extracted data. It is tightly connected to the **data quality** check where completeness, correctness, consistency and timeliness of the data are considered in relation to the research questions that ought to be answered with this dataset.

The transformation phase involves converting the extracted and cleaned data into a format that is suitable for the target system. This phase can include a variety of tasks, such as data mapping, data enrichment, data aggregation, solving syntactic and semantic heterogeneity and combining or decomposing data in columns. Sometimes it is not clear which row in one system can be combined with which data in another system. Advanced entity matching techniques have to be used, including fuzzy matching, to solve this problem.

The loading phase involves loading the transformed data into the target system, such as a data warehouse or a data lake. The loading process can be done in different ways, such as bulk loading or incremental loading. The goal of the loading phase is to ensure that the data is organized and stored in a way that makes it easily accessible and usable for data analysis, reporting, or other business processes.

Data sharing can be done in several ways. We can put all the data in a **centralized platform**. This has as a consequence that the data leaves the owner and the owner has no control over what happens to the data. Sending updates to the central platform is usually a problem.

In a **federated platform**, the data stays at the owner, and only aggregated data is stored in the central node of the federated platform. Queries are sent from the central platform to the local nodes. The owner has full control over which queries are executed on the local data and can refuse to execute them.



In conclusion, ETL is a critical process for organizations to integrate data from various sources into a target system. The extraction, transformation and loading phases of ETL require careful planning, design and execution to ensure the accuracy, reliability, completeness and timeliness of the data.

SESSION IV: PROTECTING PERSONAL DATA

An Introduction to Data Privacy and Anonymization

Session IV: Protecting personal data

An Introduction to Data Privacy and Anonymization

The transition of the healthcare sector into a data-centric model necessitates a shift in focus towards the principles of data privacy and anonymization. Medical students and professionals are increasingly tasked with safeguarding sensitive patient information, highlighting the urgency of understanding and effectively applying these principles. This 3 hour session was designed to navigate this complex landscape, aiming to empower participants with the essential knowledge and practical skills to protect patient privacy effectively.

Personally Identifiable Information (PII) is data that can identify someone when used alone or with other relevant data. Two types of PII was introduced and discussed:

- **Sensitive PII**: Sensitive PII refers to specific types of personal information that, when compromised, can result in significant harm to an individual's privacy, financial security, or personal safety. While PII generally refers to any information that can be used to identify an individual, sensitive PII specifically includes data that requires extra protection due to its potential misuse or potential for harm.
- **Non-sensitive PII**: Non-sensitive PII refers to personal information that, on its own, does not pose a significant risk of harm or potential misuse if it were to be compromised. While this information can still identify an individual, it is generally considered less critical or potentially harmful compared to sensitive PII.

Principles of the GDPR

An integral part of this course involved understanding the legal landscape that governs data privacy. The **General Data Protection Regulation (GDPR)** is a key regulation that protects data subjects' rights and outlines the responsibilities of data controllers and processors.

Data Controller: Determines the purposes and means of the data processing and is responsible for establishing the ethical and legal framework that protects personal data.

Data Processor: Carries out specific tasks on behalf of the controller. The controller must ensure the processor is authorized to perform such tasks.

The 7 principles of the GDPR are:

1. **Lawfulness, fairness, and transparency** in data processing.
2. **Purpose limitations**: Collect data only for specified, legitimate purposes.
3. **Data minimisation**: Limit collected data to what's necessary.
4. **Accuracy**: Keep personal data accurate and up-to-date.
5. **Storage limitation**: Retain data for a limited time.
6. **Integrity and confidentiality**: Protect data with appropriate measures.
7. **Accountability**: Demonstrate compliance and safeguard personal data.

Anonymisation versus pseudonymisation

Pseudonymisation:

- is a technique used to protect personal data by replacing identifiable information with pseudonyms, or artificial identifiers
- can reduce the risks associated with data processing, while still allowing for the data to be used for legitimate purposes such as research or analytics.
- Reversible in most cases

Anonymisation:

- is a permanent process that cannot be undone. Once data is anonymized, it cannot be reverted back to its original state.
- The goal of anonymizing data is to remove all personal identifiers, including indirect identifiers, to ensure that the individual cannot be identified.
- is often used in research studies to protect the privacy of study participants while still allowing for the analysis of the data.
- In RWD, true anonymization does not exist.

De-identification strategies

Several de-identification strategies were introduced and explained during our session. Only a few of them are highlighted here in this summary.

Data Suppression: This strategy involves removing or suppressing certain data fields or values from a dataset to prevent the identification of individuals. This can be done by excluding specific variables or deleting records containing sensitive information.

Name	Zipcode	Age	Sex	Disease
Alice S.	47677	29	F	Ovarian Cancer
Betty Q.	47602	22	F	Ovarian Cancer
Charles D.	47678	27	M	Prostate Cancer
David E.	47905	43	M	Flu
Emily J.	47909	52	F	Heart Disease
Fred K.	47906	47	M	Heart Disease

Original data

Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

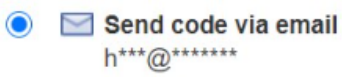
After suppression

Data Masking: Data masking involves altering or replacing sensitive data with fictional or modified values while preserving the data's format. This allows the data to be used for analysis or testing purposes without revealing personally identifiable information.

	name	phone
0	Cassandra Nelson	4399406975395
1	Brian Moss	0389407128613
2	Melody Gill	8283308773967
3	Sandra Huber	4366608954250
4	Patricia Webster	4466462475574

	name	phone
0	xxxx	4399406975395
1	xxxx	0389407128613
2	xxxx	8283308773967
3	xxxx	4366608954250
4	xxxx	4466462475574

Partial Data Masking: Similar to data masking, partial data masking involves obscuring only a portion of the sensitive data, reducing the risk of re-identification. This approach helps balance the need for data utility and privacy protection.

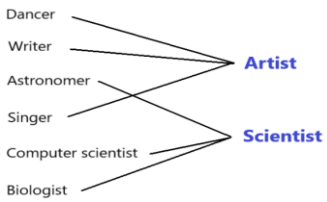


Data Generalization: Data generalization involves replacing specific values with more generalized categories or ranges. For example, replacing exact ages with age groups or substituting precise locations with broader regions. This strategy protects individual identities while maintaining data usefulness.

	first name	last name	age	ssn
0	Amber	Brown	91	798-29-4785
1	William	Gibson	34	431-66-8381
2	Daniel	Lee	92	825-91-5550
3	Andrea	Stevenson	64	188-59-3544
4	Julie	Horn	35	020-60-6388

	First name	Last name	Age	SSN
0	Amber	Brown	(80, 99]	798-**-****
1	William	Gibson	(30, 50]	431-**-****
2	Daniel	Lee	(80, 99]	825-**-****
3	Andrea	Stevenson	(60, 80]	188-**-****
4	Julie	Horn	(30, 50]	020-**-****

Data Aggregation: Data aggregation involves combining and summarizing data to produce statistical or summary information. Aggregating data helps protect privacy by preventing the identification of individuals while still providing meaningful insights and analysis at a higher level.



SESSION V: TACKLING DATA HETEROGENEITY

Understanding the importance of data harmonization and standards

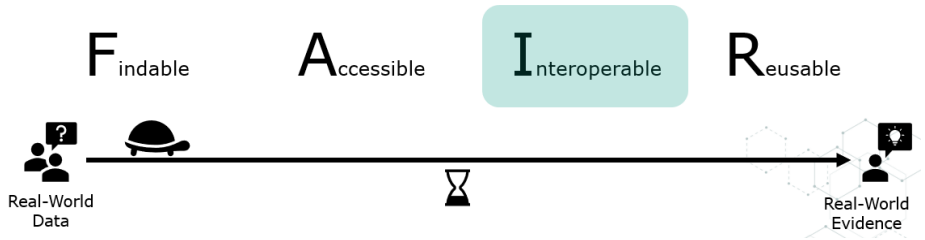





Session V: Tackling Data Heterogeneity

Understanding the importance of data harmonization and standards

We began this session by exploring one of the challenges in using RWD in health and care to generate real-world evidence (RWE): the slowness in transforming RWD into RWE. The FAIR principles, introduced for the first time in Session II, aim to speed up the time it takes to go from RWD to RWE by establishing methods to overcome bottlenecks in the transformation of RWD to RWE. One of those principles is "**Interoperability**" which acts as a facilitator of transforming an heterogeneous, disconnected network of islands into a homogeneous, connected landscape.



When looking at RWD in health and care, heterogeneity, or diversity, affects all aspects of the data and its sources. This heterogeneity complicates or even hinders collaboration between data partners with their respective data and slows down the progress in generating RWE from RWD in health and care.

Example

If datasource A collects data on topic X and data source B collects data on topic X, we would assume that a collaboration or joint analysis is easily possible. Most of the time though, datasource A uses their own defined dataset in their specific data collection tool where data is collected and stored in the A-manner, while data source B has a B-specific data representation. An easy example is the data on gender. Even if both data sources reside in one country (and may speak the same language), the representation of such an obvious variable might differ. A uses "M" and "F", B uses "1" and "0".

The ultimate goal to improve communication between data sources and promote understanding of RW datasets is to establish harmonization in the data by finding a **common language for the data.**

We differentiate between prospective and retrospective harmonization.

- **Prospective harmonization**: Establish or use a common approach to collect health data, e.g. via the use of standards in electronic healthcare records for the ongoing and future data collection.
  The way data is collected is changed

- **Retrospective harmonization:** Establish or use a common approach to assemble health data, e.g. via the use of a common data model for the transformation of existing data collections.
  The way data is collected is not changed.

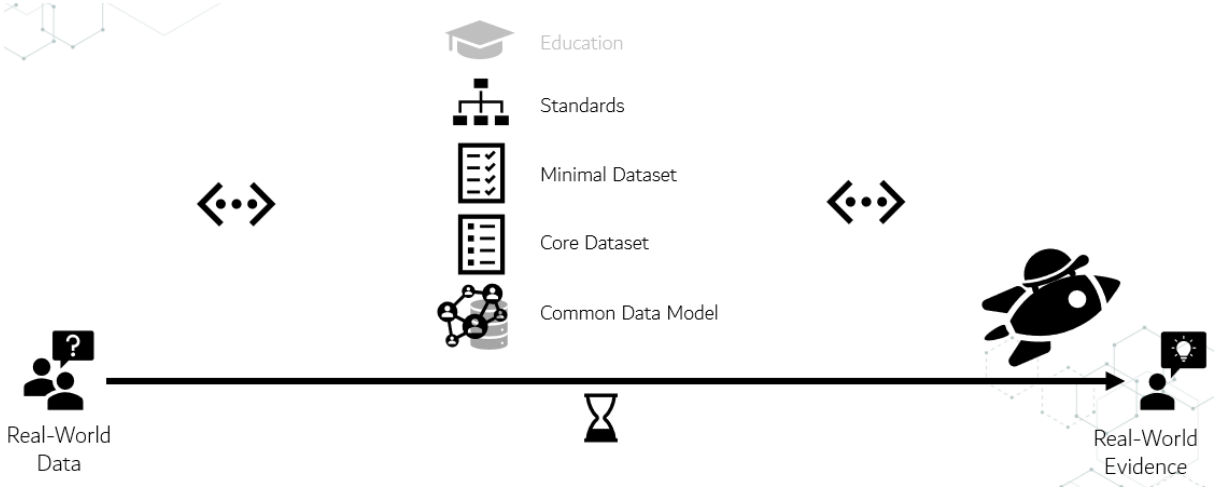
Different strategies to promote harmonization were introduced:

- **Standards**: A standard is an agreed-upon data capture and sharing format that states how data is collected, described or stored and how it can be retrieved or used to share insights. Standards in a hospital cover transfer/exchange standards, security protocols and content standards. Content standards include standardized codes and terms for data in health and care (e.g. hospital). Examples are: ICD-10, LOINC, ATC and SNOMED CT. *Standards are intended to promote prospective harmonization, but are often applied for retrospective harmonization.*

- **Minimal and core datasets:**
    - Minimal datasets: Standardized list of variables (and details) that was agreed upon within an initiative to serve as the base for a collaboration that may be limited in time, purpose or membership. The list is initiative specific, with a tailored data collection to fit the concrete initiative's needs and not necessarily globally applicable.
    - Core datasets: Can be seen as the basic ("core") data items in a specific area of interest, e.g. a certain disease, that represent the common denominator across an area of interest and the accompanying (minimal) datasets.
    - *Minimal and core datasets are mostly used for retrospective harmonization, but can also serve prospective harmonization, if implemented within the routine data collection software (as early as possible).*

- **Common data model**: Standardized representation of content, combined with a standardized infrastructure to store the data, independent from a research question or defined minimal / core datasets. Its purpose is to enable collaborative analyses by providing a defined framework and the "common language". An important example of a common data model is the OMOP common data model. *A common data model aims at retrospective harmonization, by providing standardized representation and structure of data as a target format for retrospective harmonization efforts*.

These methods of promoting harmonization are one important part in shortening the time it takes to gain real-world evidence, paving the way for a rocket-fast RWD to RWE transformation.



SESSION VI: GOVERNANCE

Effective Governance and Contract Management

Session VI: Governance

Effective Governance and Contract Management

Governance encompasses "the system" by which a project is controlled, by which a project operates, and the mechanisms by which it, and its people, are held to account. Ethics, risk management, compliance and administration are all elements of governance.

As mentioned and introduced during Session IV, one of the key principles of the GDPR is "lawfulness". There are 6 legal bases:

1. Consent
2. Public interest
3. Legitimate interest
4. Legal obligation
5. Execution of an agreement
6. Vital interest

At the university, we only use consent and public interest as a legal base. The most important differences between 'consent' versus 'public interest" as legal bases are summarized in the table below. Defining what legal bases apply for which project is non-trivial and should be discussed and checked with your data protection officer (DPO).

Informed Consent	Public Interest
1. Every data subject has to give explicit **consent to the processing** of their personal data 2. If subject withdraws consent: processing of data already collected can no longer take place 3. The informed consents of all individual patients have to be stored and managed at the level of the controller	1. If the processing is necessary for the **fulfillment of a task of public interest** 2. Urgent social need for the processing of certain personal data → implicit increase in knowledge in the interest of society 3. Effective task of public interest is assigned to the controller. 4. Data subjects have the right to know what happens with their data (transparency) 5. Cannot be used by everyone. E.g. pharmaceutical companies can't rely on this legal ground. Task must me laid down in the national law of a Member State.

Informed Consent - ETHICS	Informed Consent – LEGAL GROUND
1. A person gives consent to **participate in research** 2. Data subject needs to understand what the research is and what they are consenting to 3. Consent needs to be given prior to their participation in the research 4. Written or oral; researcher needs to store the consents 5. There are 2 stages in a standard consent process: 1. Giving information 2. Obtaining consent	1. A data subject gives consent to the **processing of the personal data** (in the study personal data is collected and processed) 2. Consent is given to each individual processing activity before the actual data collection 3. Purpose of data processing is not always clear at time of data collection ☐☐ "Broad consent" for certain areas of scientific research
• Ethical consent is not necessarily subject to the same conditions as the consent as legal basis in the GDPR. • Whenever the research entails the collection/processing of personal data (with consent as legal ground), one form and information letter may cover both the ethical and the legal consent. • Consent can be withdrawn at any time.	

As an example of a governance and use pipeline, we introduced the pipeline that is currently followed by the research group in Biomedical Data Sciences for research projects where patient-level data is reused in a centralized approach (data is transferred). Before starting the project in which personal and sensitive data is reused, one needs to think about which variables will be needed, the methods of the research, protocol that will be followed, where the data will be stored and for how long, what the impact of the data processing will be, how risks can be limited,... ("Privacy by design"). A data management plan (DMP) and/or a concept note containing all the specifics of a project could help with this.

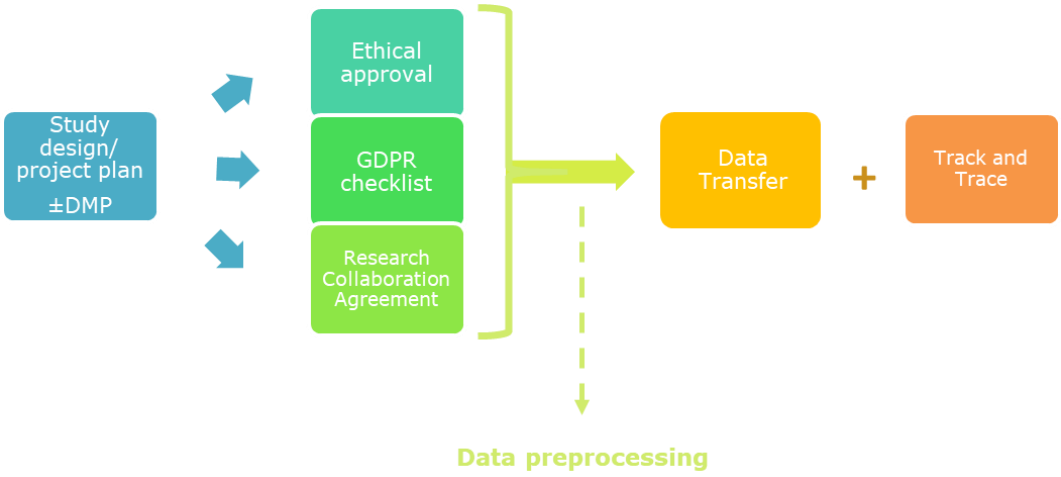


Figure 1: The pipeline which is followed in the research group of Biomedical Data Sciences when personal data collected during standard care is transferred in the context of research.

Several administrative steps need to be completed before the data can be transferred:

- **Ethical approval** needs to be granted by the ethical committees of all institutes involved
  - *For retrospective studies (reuse data that is already collected), you don't need informed consent from the patients.*
- **GDPR checklist** (UHasselt-specific) needs to be completed
  - *Data processing is based on a legal ground; most common legal grounds used in research projects at UHasselt are Consent and Public Interest. Important note: Informed Consent (IC) as lawful basis ≠ the IC that might be needed for ethical approval*
- A **contract** between all parties involved needs to be set up
  - there are different kinds of contracts, e.g. research collaboration agreement, service contract,...

Because of the GDPR principles of data minimization and purpose limitation, only the data that will actually be needed to perform the research can be transferred. Therefore, datasets are often preprocessed in order to downsize it to the extent that's really needed for the research.

Although the high-level process is always the same, every project and every dataset is different and therefore, it's not always easy to decide on governance-related matters. Context and specifics about projects are determining how and by whom certain paperwork needs to be completed. However, you're not expected to decide everything on your own. Students and employees of UHasselt can and should ask help or assistance from the data protection officer(s), business developers, tech transfer office and research data management team of UHasselt.

SESSION VII: FROM DATA TO INSIGHT

Generating data-driven insights to address urgent questions






Session VII: From Data to Insight

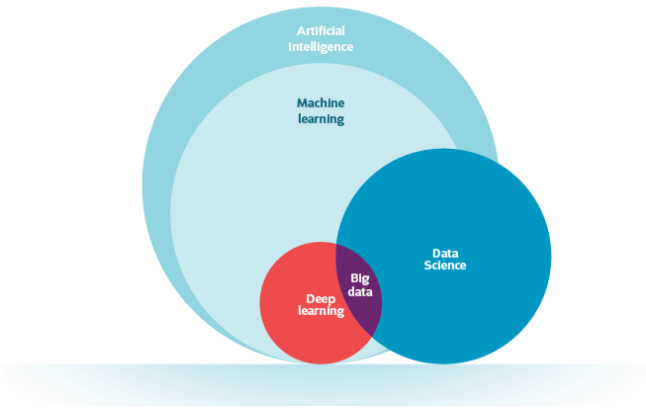
Generating data-driven insights to address urgent questions

Patients leave a trail of data through the health system, at their general practitioners, the pharmacy, insurance companies or within the numerous hospitals' IT systems. This vast amount of data has the potential to drive clinical decision-making but poses a challenge before gaining insight: it is nearly impossible to inspect such volumes of data manually. Automated methods to gain insights into large amounts of data are needed and have been developed for centuries. This session consisted of three topics that aim to generate insight from data. Statistics represent the simplest but most mature form of generating insights into data. Machine learning algorithms, often seen as a direct descendant of statistical methods, bear a larger potential to identify data patterns and thus gain insight. Finally, deep learning approaches have shown to generate innovative insights with the downside of necessitating vast data collections.

Gaining insights with statistics

Descriptive statistics are basic, well-known concepts taught in school. Describing a collection of data with intuitive key figures like averages, standard deviations, or correlations is part of the basics of math. Albeit basic, those key figures still drive medical decisions and are integral to evidence-based medicine. A handful of statistics only tell a part of the story, so we use graphs to visualize data. Typical examples include bar plots, scatter plots or line plots.

Artificial intelligence basics



There is a massive push in the healthcare industry and medical research to employ artificial intelligence methods to improve patient health and care. However, *artificial intelligence* is a philosophical term that is not well-defined. We typically refer to either *machine learning* or *deep learning* when talking about artificial intelligence. Machine learning describes an approach to performing specific tasks with a computer without explicitly programming it. A subset of such approaches, deep learning, represents a particular set of algorithms that mimic the brain's structure to solve specific tasks.

Machine learning: data, training, and modeling

Machine learning approaches, also named *models*, typically follow three phases: *training*, *testing* and *use*. During the training phase, a model will repeatedly see some data entries to learn its structure or its content. As data is the basis for each machine learning model, it will determine which type of machine learning models we can use.

- *Supervised learning* describes training on a dataset labeled with the desired output. For each data input, something or someone documented the result. We consider models based on supervised learning to be *task-driven*.

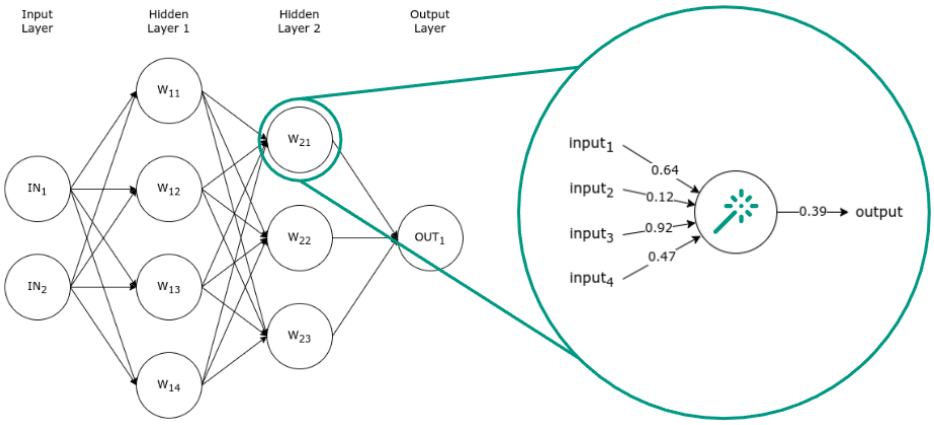
- In contrast, *unsupervised learning* describes training a model on a dataset without labels. Hence, an unsupervised model has to learn to generate the desired output based on the structure of the data itself. We consider unsupervised models to be *data-driven*.

- As a third option, we consider *reinforcement learning*. With the help of an interactive environment, a model retrieves direct feedback on its output and therefore *learns from its mistakes*.

We aim to evaluate how well a model works in the testing phase. We do this typically by withholding some of our training data to evaluate a model on data unavailable during training. We call the different parts of the data *splits*. The *training split* is used to train a machine learning model. We use an optional *validation split* to test various settings on a single model. Finally, the *test split* is used to assess the performance of a machine learning model. The testing phase allows us to avoid the two most common pitfalls of machine learning. If a model is too specific for a single dataset, we call it *overfitting.* Such models will perform poorly on unseen data. As opposed to overfitting, we call *underfitting* a training process that was too generic, performing poorly even on the training data.

Deep learning: neurons, neural networks and learning

At the basis of deep learning approaches, which is a subset of machine learning, is the *artificial neuron*. Modeled after their real-life counterparts found in mammal brains, they take some (numerical) inputs, assign a weight to each input and calculate a (numerical) output. On its own, this rather abstract idea has no use. Linking several neurons together in a particular structure, where we usually build them layer by layer, unfolds an immense potential that has proven itself in various use cases. We call those structures *neural networks*. Mathematically speaking, a neural network is a collection of weights and a description of how to apply them consecutively to the input. To create an effective neural network that can solve a task, we train it using two concepts: *forward propagation* and *backpropagation*. We take a single instance of input data during forward propagation and run it through the network. After this, we apply backpropagation. We compare the obtained output to the expected output and calculate the difference using the *loss function*. We traverse the network backwards, from output to input, and calculate the loss at each neuron. The computed loss gives us the possibility to update all weights. If we repeat this process with enough data, the neural network learns to give the correct output over time.



SESSION VIII: TRUST AND BIAS/ETHICS

Finding the Balance between Data Protection and Data Benefit

Session VIII: Trust and Bias/Ethics

Finding the Balance between Data Protection and Data Benefit

Prof. Dipak Kalra is the president of **The European Institute for Innovation through Health Data** (i~HD: https://www.i-hd.eu/). The institute's mission is to bring together diverse stakeholders across the healthcare and clinical research spectrum to help co-create solutions we need nowadays for:

- the capture and share of better quality health data;
- its trustworthy use for smarter health care and efficient research.

We can think of opportunities for using health data at the **individual, population, or big data level.** The opportunities for the (re)use of data range among others from individual health status monitoring and personalized medicine to population-based pharmacovigilance and big data-based epidemiology, digital innovation, and AI- and drug development.

In Europe, many initiatives of data infrastructures to share health data across jurisdictional, institutional, and domain borders are arising for direct patient care and research purposes (e.g., MyHealth@EU, EHDEN, Elixir, Health data Hub, EU RD Platform, etc.). The most exciting current initiative is **the European Health Data Space** which is to be a data ecosystem connecting all 27 EU member states for both continuous patient care across countries and multi-country data access for research. There is a draft regulation currently going through the European Parliament. We are optimistic that this proposal will result in an infrastructure that will allow for the many ways in which data needs to be connected across Europe for research purposes and a set of governance rules that make sure that citizens are informed and able to make decisions about their data flows and uses.

Navigating the Ambiguities of Data Protection Law

The data controller must adhere to the GDPR principles, as previously depicted in Session IV. However, there are a number of **challenges associated with the GDPR and using big data** for research within the context of emerging data infrastructures:

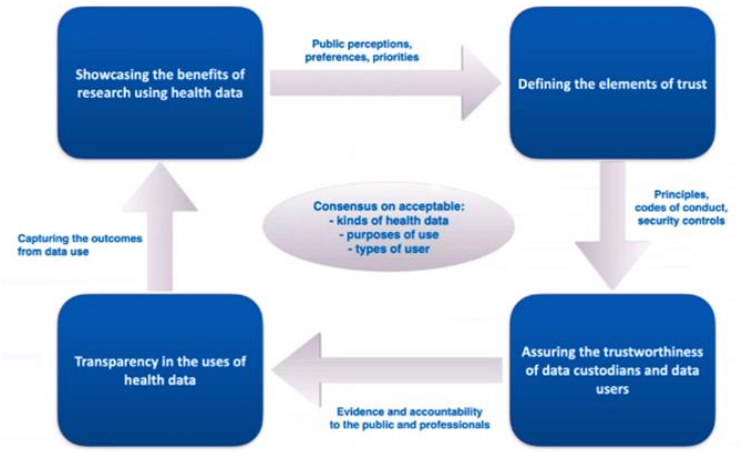
1. **Precise consent explanation**: The GDPR mandates that data must be held for a specific purpose with a lawful basis, usually consent. However, explaining to patients how their data will be used for multiple purposes and how precise these explanations should be is challenging.

2. **Realistic consent requirements:** Obtaining consent from multi-million patients every time data is used in a big data infrastructure is impractical. The question of whether consent is realistic each time data is reused needs to be addressed.

3. **Data minimisation challenge:** When planning for multiple data reuses, limiting the dataset to only the data required for a specific purpose is problematic. It is difficult to be minimal with the data held when there is an open-mindedness about future purposes that may require additional data.

4. **Anonymised data updating:** Retaining one-way linkage with anonymized data for updates is challenging, particularly in the case of health data, where certain identifiers are necessary to match updated information with old data. The question of whether anonymized data can retain enough identifiers, classified as personal data under GDPR, for future updates is a concern.

Building a Trust Ecosystem: A Holistic Approach is Needed



To build public trust in the use of health data, a holistic approach is required that involves starting with a consensus on which types of health data are acceptable to use, for what purposes, and which credentials of the user should be considered trustworthy. **Examples of past research benefits** from using health data are needed to engage with society, **understand their preferences and priorities**, and factor in the elements of trust to determine areas of assurance. **Principles, codes of conduct, and security controls** should be distilled to deliver against the trust and provide assurance to society. Both data custodians and users must be trustworthy in collecting, processing, using, and safeguarding data. **Collecting evidence of trustworthy use** is essential, and clinical professionals who often collect the data must be confident that it will be used in a trustworthy way. The public must have access to information on how their data is used and what benefits arise from it. Implementing this cycle can accelerate trust and the use of health data, reinforcing its ability to benefit society.

Will the Societal Compact for health data reuse be widely adopted by stakeholders?

A proposal for a **Societal Compact** for the secondary use of health data was developed by a multi-stakeholder expert group in late 2022 (led by the Digital Health Society (DHS: https://thedigitalhealthsociety.com/) and I~HD), and it is currently in draft form for a consultation. The purpose of the Compact is to **provide assurance to all stakeholders, particularly the public,** that health data will be reused legally, ethically, and securely in the interest of society.

The Compact consists of several components, including ethical principles, the scope of applicability, permitted and prohibited purposes, data use commitments, governance rules, the adaptation declaration, and templates. Once adopted, the success of the Societal Compact, which provides society with a standardized set of promises for the secondary use of health data, will depend on **two key players: data user organizations and the public,** with the former needing to sign up for the promises and declare them reasonable while the latter must be assured that the Compact is detailed enough to address their concerns about the use of health data, especially when it includes large volumes and numbers of data uses too complex to explain.

5. **Challenging data types:** Certain types of data, such as genomics and images, are difficult to anonymize, given that every individual has a unique pattern.

6. **Interpretation of adequate safeguards:** The GDPR requires that safeguards and security measures taken for anonymizing data be "adequate" and "proportionate." However, the interpretation of these terms varies across different member states, making it difficult for researchers to obtain a consistent plan for accessing and protecting data across multiple locations.

7. **Data user expectations:** It is unclear what promises can be expected from a data user, and there is a lack of coherence across Europe regarding what constitutes reasonable expectations and how to enforce them. It is challenging to ensure that data users uphold these expectations and address issues of misbehavior.

8. **Consistency across Europe:** Creating consistency across the 27 EU member states is necessary to avoid a patchwork of different approaches, making research impractical.

What Hinders Public Acceptance of Data Reuse in Health Research?

Society is nervous about their personal data being accessed for research purposes, particularly by the industry. However, research by the industry is needed to innovate and deliver new solutions to manage complex diseases. The public acceptance of health data reuse decreases as data moves from individual to the population to big health data levels. **Why does society struggle to trust big data research?**

1. There is a **lack of understanding** of how and why big data is used.
2. People are also **unfamiliar with the companies**, actors, or organizations using big data.
3. **It takes a long time** to go from the use of big data to its value back to society.
4. The research results **might not even benefit** the particular individual patient.
5. People feel they have **little control** over these distant data uses and see cybersecurity as a greater risk.

However, by making an effort to explain how data can be used safely and usefully, a high proportion of people are willing to support research uses. The public needs greater transparency about why and how health data are used, safeguarded and the benefits of that use. To reach societal acceptability, we must find a balance between individual rights and societal benefits.

SESSION IX: TOWARDS IMPACT

Implementation of cutting-edge technology in real-world practice

Session IX: Towards Impact

Implementation of cutting-edge technology in real-world practice

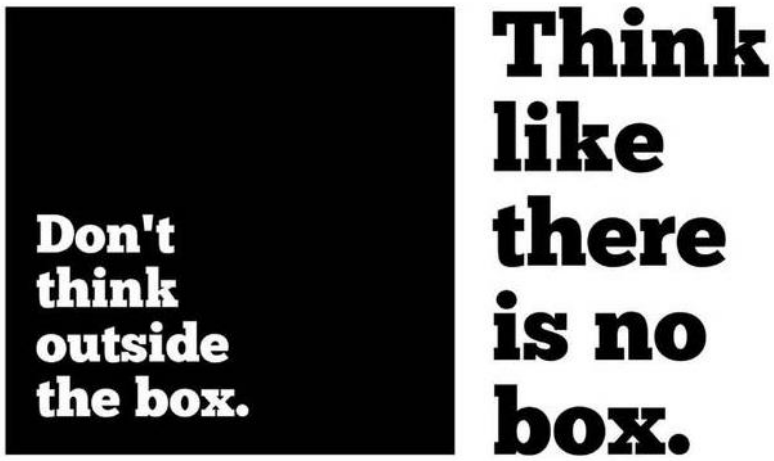
Implementation science can be defined as "*the scientific study of methods to promote the systematic uptake of research findings and other evidence bases practices into routine practice, and hence, to improve the quality and effectiveness of health services*". Research in quality and safety is an essential field that aims to improve patient outcomes, reduce medical errors, and enhance the overall quality of healthcare. If you are interested in working in healthcare, this topic is a must-know. Overall, the goals of research in quality and safety are focused on improving patient safety, reducing medical errors, and improving the overall quality of care. By identifying risk factors, developing, and testing interventions, and evaluating the effectiveness of interventions, researchers can help healthcare organizations improve patient safety and healthcare outcomes. Several approaches to research in quality and safety exist. Knowledge translation focuses on translating research evidence into practical applications, while implementation science aims to study the best methods for integrating research findings into practice. Improvement science, meanwhile, is focused on identifying areas for improvement and testing and implementing changes to achieve better outcomes.

Table 1 Comparison between approaches on patient safety research

	Knowledge translation	**Implementation science**	**Improvement science**
Starts with	Evidence-based research findings	Innovation	Quality & safety concerns
Main question	How to bring evidence-based research findings into clinical practice	How to implement innovation in practice	How to define, measure, and improve quality and safety
Definition	The 'exchange', 'synthesis', and ethically sound 'application' of researcher findings with a complex system.	The scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice, and, hence, to improve the quality and effectiveness of health services.	The scientific study of methods to systematically measure, monitor, and improve quality and safety of health services.
Audience	Healthcare professional Policy makers	Healthcare professional Healthcare organisations Public health	Healthcare organisations Healthcare systems

Tools	Systematic reviews Meta-analysis Evidence-based practice Journal clubs	Theories, models, and frameworks Behaviour change interventions Nudging	Q&S indicators Process innovation Systems innovation
My simple view	What to implement	How to implement	Why/what and how to improve

All three approaches are important for promoting evidence-based practice and improving healthcare delivery and outcomes. In this session you were introduced to the basics of research in patient safety and quality improvement. One of the most exciting aspects of research in quality and safety is the potential for innovation and improvement. From developing new technologies to implementing new care models, researchers in this field are constantly pushing the boundaries of what is possible in healthcare.



The THINK³ simulation & innovation lab is a university civic lab for education, research, and service in the healthcare sector. It provides an interactive and innovative learning environment for students, teachers, and professionals in the field, where transdisciplinary research and learning in healthcare are central. The goal of the THINK³ lab is to support (healthcare) organizations with these challenges. THINK³ aims to be a bridge between the university and the field, as well as between education, research, and service. This is achieved by actively engaging in skill training, developing innovative ideas, and implementing them sustainably.

The name THINK³ was chosen to represent the vision where three ways of thinking are central:

1. strategic thinking
2. critical thinking
3. and design thinking.

These modes of thinking are essential for appropriately assessing long-term developments in healthcare and well-being, understanding complex issues, and developing and implementing workable solutions in practice. The THINK³ simulation & innovation lab is an open lab where you can enter in an easily accessible way with different challenge(s) in the field of (health) innovation.

More information is provided in a recent discussion paper that can be consulted using this link.

SESSION X: LEADING CHANGE IN A COMPLEX ENVIRONMENT

The role of co-creative leadership and multi-stakeholder collaboration

Session X: Leading Change in a Complex Environment

The role of co-creative leadership and multi-stakeholder collaboration

People have no trouble with change as such, and certainly not if the change is aligned with their own aspirations. However, people don't like to be changed. They have their own perceptions, interpretations, interests, fears, energy levels. That's why the soft, intangible part of change is often the hardest to deal with. Leading change is in this regard so much more than managing change. Managing change is the easiest part.

Building a guiding coalition

"Never doubt that a small group of thoughtful, committed, citizens can change the world. Indeed, it is the only thing that ever has." Margaret Mead probably was referring to Kotter's idea that the establishment of a guided coalition is an important condition of successful change. What characterizes such a small group as committed citizens or guided coalition?

- Its members deeply share the same aspiration or WHY of the change.
- They share the ambition to give the envisioned change PRIORITY over other pursuits, projects and activities (versus 'management by decibels').
- A guided coalition has a heterogeneous composition: its members represent all stakeholders who have an impact on the change in terms of support, resources and execution.

A guided coalition is a group, which is much more than just a bunch of people sitting at the same conference table. In mature groups self-interest is transcended, political games are dismantled, and there is a sense of trust and psychological safety to learn in the process of collaboration and from the mistakes that inevitably will be made.

Dealing with power differences

When multiple parties are invited in the room, each of them brings along their own knowledge system (experiential versus theoretical, medical, legal, IT, …), their own expert language, their own interests, and their own framing of the problem. How can we connect in the face of this diversity? For instance, between those who want change (on a global/strategic level) and those who will have to change (on local/operational level)? The issue of diversity is closely related to the issue of power. When people meet power comes into play. Sources of power are societal status, educational, financial, moral, …. It is very important not to neglect power dynamics. Here, five leverages are mentioned briefly.

1. Who's invited at the table? Who is allowed to participate? Organizing shared activities apart from the 'official' meeting can be helpful to reduce power difference.
2. Who is allowed to co-define the problem? An important distinction can be made between participation and collaboration. In a participatory process , it is the leading actor who defines the problem. In a collaborative process there is a search for a joint problem definition.
3. What conversational design can facilitate the exchange? Examples are in-group preparation time, bilateral meetings, multilateral meetings (*cfr.* fish bowl); equal speaking time, talking stick, inviting low power parties to speak first, etc…
4. Facilitation of the ongoing interaction in the here and now in two ways. (1) By supporting interventions of low-power people (Inviting them to speak first or to speak when they remain silent.) (2) By addressing the nonverbal communication of high-status people: "I see you

frown. What does that mean?" "I notice you interrupted person X." In doing these kinds of interventions, low rank parties can be empowered to express their voice, and high rank people can be asked to listen more accurately.

5. How can the attendees be invited to reflect on their own patterns of interaction and to communicate about how the communication is perceived? This meta-communication can also increase the awareness of their mutual interdependence.

Co-creation a change vision

What is the guided coalition's first task? Co-creating a shared change vision. This vision deals with three issues:

- Envisioning an inspiring 'end state', that reflects the stakeholder's perspectives and their shared interest.
- Defining the milestones which consider the actual level of demands and pressure on the change recipients. Defining the 'moments of truth': *i.e.* "How are we going to notice (symbolically) that we are changing?"
- How are we going to communicate our change vision in a way that reflects the perspective of the change recipients and that takes into account prior change successes and failures that color the change recipients' expectations? Communication needs to be (1) honest and transparent, also regarding the expected difficulties, and (2) brought in an empathic and supportive manner.

Empowering action, co-creation and motivation

How are people motivated to change? At the point where the change vision is communicated others need to become convinced that the change is needed and possible. The metaphor of the rider sitting on an elephant walking on a path.

- The rider stands for our rational brain that wants to know the facts and wants to understand what the change is all about. Management feeds our rational brain. However, when the rider wants to go in a certain direction, but the elephant doesn't feel like it, it is obvious that the goal will not be reached.
- The elephant refers to our emotional brain: people will only change if they feel the need to change. Here, the importance of an experiential (rather than merely an intellectual) understanding of the change and positive feeling about the change. This is where leadership comes in.
- The availability of a path can make the journey easier. So, here the question is how the environment can be tweaked, how favorable circumstances can be created, what enabling tools can be developed, or what knowledge can be made available.

So, people will be more likely to go along when their brains intellectually understand the change, when their hearts feel like changing, and when the circumstances invites their feet to walk the path. Once the direction is set by the guided coalition, other people are invited to engage based on their perspective, knowledge and motivation. Where they are heading is clear, but how to get there needs to co-construct or co-created by all people involved. In doing so they become empowered.

Dealing with resistance

Where there is power, there is resistance. (Michel Foucault). Where there is terrorism, there is imperialism. (Arnold Mindell). Dealing with resistance is done in three steps:

1. Noticing resistance. Resistance can be expressed in very different ways. Often it is also implicitly or covertly. That's why it is important to bring resistance to the surface by encouraging people to raise their concerns. For instance, by asking questions such as: "who does not agree?", "who is unhappy with the way things are going?" Observing and addressing nonverbal behavior can also be very helpful.
2. Exploring resistance means that you try to understand why, by asking questions such as: "which of your interests, needs … are not being met?" Resistance can have many sources.
3. Transforming resistance by showing empathic understanding (step 2 which can lead to grieving and dissolution of the resistance) and by finding solutions that take into account the concerns that people have.

Personal leadership

Finally, there is an important sixth theme, which we will not go into at length here. Change can be quite a challenge. And when we are challenged, our personal leadership is put to the test. At moments we will feel discouraged, irritated, impatient, frustrated, … (mindstate). And there will also be moments in which we are caught in judgments and wrong assumptions (mindset). How can we relate to ourselves in a way that we can cultivate more generative mind states and mindsets?

Change - such as the #DataSavesLives-movement - is "world work". And change also requires "inner work". This dimension is often overlooked, and its importance is often underestimated.

RESEARCH GROUP BIOMEDICAL DATA SCIENCES

Visit our website: