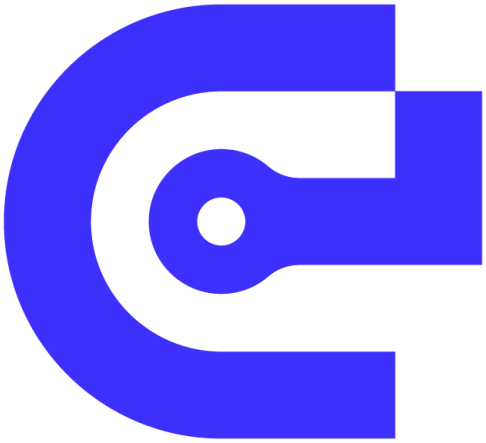
Co-funded by the
Erasmus+ Programme
of the European Union



Cooperation and training on innovation and entrepreneurship in the eHealth community (CONNECT)

2020-1-RO01-KA203-080244

IO1 - eHealth Interdisciplinary Curriculum: Health Analytics and Big Data in Health

Co-funded by the
Erasmus+ Programme
of the European Union

Partner: University of Porto

Porto, 30th September 2021

eHealth Interdisciplinary Curriculum Template – Purpose of this tool

Babeș-Bolyai University has developed this tool as a guide and generic template for creating the eHealth Interdisciplinary Curriculum. We have tried to make it user-friendly by providing explanations and examples under each heading.

The eHealth Interdisciplinary Curriculum will be developed under *"Objective 1: Develop an innovative multidisciplinary curriculum for students from the computer and information, healthcare and social professional background, with the main focus on cooperation between sectors for improving the existing knowledge, skills, and accessibility to new opportunities"*. The indicators of this objectives are represented by 1 curriculum developed in the first 10 months of the project, with at least 1 member of each partner institution involved in the curricula development.

The eHealth Interdisciplinary Curriculum is centered around theoretical and practical subjects within the eHealth domain. It will have the form of an online book, adapted as an interactive online resource, and uploaded on the online platform for managing eHealth eLearning. It will be addressed to health sciences and IT students, from participant countries and disseminated to students from other European universities. This Curriculum will focus on undergraduate students, but other beneficiaries can be included. Although there is a requirement that readers and learners need to have a background in health care/ medicine/ information technology, information systems or business.

The eHealth Interdisciplinary Curriculum will include foundational knowledge (formal), key perspectives in eHealth (examples of new technologies, applications, instruments – non-formal), application abilities (increasing qualifications, competencies, and critical thinking – non-formal) to

provide eHealth remedial education. Consultation of formal and informal educational providers will be necessary in developing the curriculum.

The eHealth Interdisciplinary Curriculum is organized to emphasize relationships between different fields (health, IT, management). It will be structured on the recommendations of the International Medical Informatics Association (IMIA).

The primary learning goals of the curriculum will be integrated to create a coherent methodology: (a) foundational knowledge (concepts, principles, facts, terms), (b) key perspectives in eHealth, that will be the starting base of practical abilities, (c) application abilities - to have a standard of working competencies for the future workplace, (d) to engage students to increase interest and have access to information.

The eHealth Interdisciplinary Curriculum will be developed by an international, inter-professional teaching team (members) with different expertise in the eHealth domain, from partner institutions. Two educational providers, from each partner institution, will be involved in the process. For each chapter, at least two external contributors will be invited to co-author the chapters and give feedback on the developed intellectual output.

The eHealth Interdisciplinary Curriculum will be purposefully designed (flexible, modular format, user guidance) so that they can be easily used and transferred in academic activities and within the university curriculum. The eHealth Interdisciplinary Curriculum is comprised of 8 individual modules. The number of pages of the entire Curriculum will be between 200-300, A4 format– around 30-40 pages/module. The course material for the entire Curriculum requires 40 hours of the hands-on, active reading experience. For each module a maximum of 5 lessons plans of 1 hour each are recommended (5 hours/module). Extra 20 hours must be added (for necessary time to access references and areas of inquiries) for the entire Curriculum, meaning 30 minutes for each lesson plan (2.5 hours for each module).

The following steps will be taken for the development of the eHealth Interdisciplinary Curriculum:

1. Desk Research
2. First draft developed by each institution for their module
3. Expert review and input

4. Second draft developed by each institution for their module based on the expert input
5. BBU complies final version of the curriculum
6. Experts validate the final curriculum

The research team from Babeș-Bolyai University is available to support any efforts to compile each curriculum component (module) and is responsible for overseeing the compilation of the final eHealth Interdisciplinary Curriculum. The contact info for the coordination team for this task is provided here: madalina.coman@publichealth.ro and alina.forray@publichealth.ro. Please name the final document using the following strategy "CONNECT Project_IO1_Curriculum_Module name_Institution Acronym" (e.g. CONNECT Project_IO1_Curriculum_mHealth_BBU)

Some tips for developing the Curriculum for the assigned modules:

- Review the Desk Research documents available for all the modules and extract the appropriate information to be used for the development of the module;
- A total of 5 hours for the lesson plans and 2.5 hours for individual work are assigned to each module
- Plan for maximum 5 lesson plans, each with the duration of 1 hour + 30 additional minutes for further references and inquires that will be done individually by students;
- Describe in detail each lesson plan following the suggested headings from section *3. Lesson plans;*
- Consult the key expert points from the Expert Network Centralizer in the development of the curriculum for the assigned module.

Contents

1. Learning objectives of the Health Analytics and Big Data in Health module

This course has the general objective to instruct students on the fundamental concepts of health analytics and big data, enabling students with full comprehension and critical attitude on the complex process of current data science field of work - that is the base for decision-making in health. It is expected that students understand the process in collecting, managing and analyzing health data, as well as approaches for transposing the results of this analysis to the decision and management of health services and its application in social problems.

At the end of this course the student must:

1. Know the basic concepts and principles of health analytics and Biga Data (use of secondary data and some of its sources, data collection and management, intelligent data analysis, and incorporating these concepts to health decision)

2. Know basic concepts of health information systems (technical terms used in Health Information Systems; Information Systems architecture and processes and main sources of databases)

3. Understand and describe the main theoretical constraints on Big Data and Artificial Intelligence systems

5. Know the theoretical foundations and challenges of Data Governance and Data Policies.

2. Foundational knowledge of the Health Analytics and Big Data in Health module

Data plays a key role in modern industry and any organization. As healthcare systems continue to adopt innovative technologies for different purposes (e.g. epidemiological surveillance, monitoring, treatment or diagnostic), the volume of available data also continues to grow, both due to the increase availability of information and also the capacity to store it. (Shortliffe & Cimino, 2014) The collected amount of such large and complex data, which is - by definition - difficult to analyze and manage with traditional software or hardware, characterizes Big Data in healthcare. Nevertheless, the available information is often insufficient and limited, especially when relying on secondary data or since the available tools do not always allow the adequate collection, analysis and interpretation of data to generate quality information to apply the best interventions or decisions. In fact, the volume of data collected does not necessarily mean that it can be aggregated into useful, valid or reliable information. Transforming data into valuable information is still a challenge in health ecosystems. (Cruz-Correia et al., 2009; Gao & Yu, 2020)

Decision making can be a complex process and always involves some degree of uncertainty. Integrating individual clinical knowledge with the best available evidence from systematic investigation enhances the possibility to convert data into value and increases objectivity and confidence in taking decision to action. Through a stepwise process decision-making understands how to use and apply information to create knowledge and wisdom, allowing to increase effectiveness during the decision process. Thus, it becomes evident that health care systems and

providers have become increasingly focused on the need to use evidence to inform and make clinical and operational decisions. (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996)

The growing need for evidence-based decision-making in the clinical and governance process has revealed the need to strengthen a set of principles and practices that ensure data quality throughout its lifecycle – highlighting some data governance challenges. The objective of data governance is to ensure data lifecycle management and implementation of data quality management strategies. By integrating a set of processes that aim to implement and maintain an organizational culture of data quality which must produce, maintain, execute, and communicate data quality management practices it may be guaranteed the specifics of quality requirements and ensuring continuous data improvement. (Mehta & Pandit, 2018; Shabani, 2021) (Magnuson J.A., 2020) (Cruz-Correia et al., 2009)

The Concept of Big Data

A collection of data sets that is so large and complex that it becomes difficult to process using hand database management tools or traditional data processing applications. (Halevi, 2014) The advent of big data brings new challenges in translating datasets of various quality, quantity, and velocity (3 V´s of Big Data) into actionable information, and ultimately, to knowledge. Big data are of significant interest to the public health domain due to the size, diversity, and complexity of varied data sources that could prevent disease and promote health and wellbeing.

Big Data includes real world data such as electronic health records, registry data, claims data, data from wearable devices, social media platforms among others. Moreover, integration and analysis of the data with different nature, such as social and scientific, can lead to new knowledge and intelligence, exploring new hypothesis, identifying hidden patterns – which would be difficult (or even impossible) otherwise.

Data can often be collected in real time (e.g. monitoring patients through wearable devices) which require specific technology. In other hand, large amounts of data have already been collected for different purposes through the years. Hence, secondary data refers to data that have already been collected for some other purpose. (Schlomer & Copp, 2014) This highlights some constraints in analyzing and interpreting this amount of data - as it was not controlled for the current intended purpose.

Guaranteeing data quality through its life cycle requires a robust information system infrastructure. In contrast, a robust information system infrastructure requires the ability not only to provide and make available quality data, but also to receive data, so they need to support bi-directional communication (of alerts, population health statistics and case or care management) - to inform clinicians and decision-makers in real-time. Information architecture (AI) refers to the logical configuration of various elements, including hardware, software, information flow and technical standards needed to support the information needs of users. Robust AI can increase the effectiveness and scope of its performance by integrating internal and external information systems. A fundamental component of information architecture is interoperability. Interoperability is defined as “the ability of a system to exchange electronic health information and use information from other systems without additional effort by the user”. Problems with data interoperability (ie, send, receive, find, and eventually be able to of use) restricts data exchange with other interested parties. (Janssen & van der Voort, 2020; Magnuson J.A., 2020)

Big Data Analytics

Big data analytics covers integration of heterogeneous data, data quality control, analysis, modeling, interpretation, and validation. (Halevi, 2014) Application of big data analytics provides comprehensive knowledge discovering from the available huge amount of data. Big data analytics seeks to leverage improvements in computer science to address these needs.

Analytical approaches can be divided in three categories, namely descriptive, predictive, and prescriptive. (Magnuson J.A., 2020) Overall, Big Data Analytics can be understood as an umbrella term for data analysis applications in the context of Big Data, namely using algorithms to analyze data: regression analysis, simulation, supervised and unsupervised machine learning methods, among others. (Halevi, 2014; Watson, 2019)

Such applications of big data analytics can improve the patient-based service, to detect earlier spreading of diseases, generate new insights into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better treatment methods or increase cos-effectiveness in health interventions. (Watson, 2019) Such applications are , however, not free from limitations or disadvantages. (Pastorino et al., 2019)

Artificial Intelligence

On 8 April 2019, the High-Level Expert Group on AI at the European Commission presented Ethics Guidelines for Trustworthy Artificial Intelligence. (Comission, 2019) In this report, Artificial intelligence systems (AI) were defined as " software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

Challenges

The previous definition raises awareness on some constraints that should be considered. Any analytical process is only as good as the quality of data available for analysis. As such, it is important to ensure that datasets used for analysis are cleaned and parsed to present a concise, valid, and clear picture of the datasets being used. Moreover, trusted and reliable analytics and artificial intelligence — which are key ingredients for transparent, evidence-based policymaking— require findable, accessible, interoperable, secure and high-quality data. In fact, regarding collection of large amount of data and the application of such systems, some challenging issues should be considered. (Maissenhaelter, Woolmore, & Schlag, 2018) Therefore, latter in July 2020, parallel to the following necessity raised by the COVID-19 pandemic, it was released a strategy for data governance and data policies at the European Commission (Comission, 2020), focusing on processes thar endure (1) Data Governance and management; (2) Protection and information security; (3) Data Quality; (4) Interoperability and standards.

Proposal for the module:

		THEORIC	PRACTICE	HOURS
FUNDAMENTALS		Concepts definition		
		Big Data and secondary data		
		Data sources and types of data		
		Information systemns in Health		1
BIG DATA ANALYTICS		Introduction to Descriptive, Predictive and Prescriptive analysis		
		Computational methods for large databases (analytical and modeling techniques)		
		Artificial intelligence in healthcare: Fundamentals		
		Artificial intelligence in healthcare: Issues		1
DATA-DRIVEN DECISION MAKING		Principles and definition of Data Governance		
		Challenges of Data Governance		
		1 Data Quality		
		2 Protection and Information Security		
		3 Data Interoperability and Standards		
		4 Data Governance and Management		1
CASE STUDIES			Case 1	1
			Case 2	1
			TOTAL	5

3. Lesson plans for the Health Analytics and Big Data in Health module

Lesson plan 1: Fundamentals

Foundational knowledge

1. **Concepts definition**
    a. **Evidence-based making decision making (Hunink et al., 2014; Shortliffe & Cimino, 2014)**

Data are central to all health care as it is crucial to the process of decision**.** In fact, all health care activities involve gathering, analyzing, or using data. Data provide the basis for categorizing the problems and identify subgroups, patterns or outliers within a population of patients. Evidence-based Decision Making include healthcare policy decision making, public health and population-based decision making (in the form of guidelines using formal evidence criteria and processes). In fact, it is a process for making decisions that is grounded in the best available research evidence and informed by experiential evidence from the field and relevant contextual evidence (see example 1).

b. What is Big Data? (Halevi, 2014)

Big data means there is more of it, it comes more quickly, and comes in more forms. It is a collection of datasets that is so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications.

c. The V's of Big Data

The definitions and assessment of big data are driven by three concepts often referred to as the three V's, volume (quantity of data), variety (types of datasets), and velocity (how often the data are being captured/reported)

d. Big Data Analytics

Big Data Analytics can be understood as an umbrella term for data analysis applications in the context of Big Data, namely using different algorithms to analyze data. Big data analytics covers integration of heterogeneous data, data quality control, analysis, modeling, interpretation, and validation.

e. Big Data Technologies (Davenport & Harris, 2007)

Requires new technologies/techniques to collect, store, analyze and visualize it. Some examples are provided below:

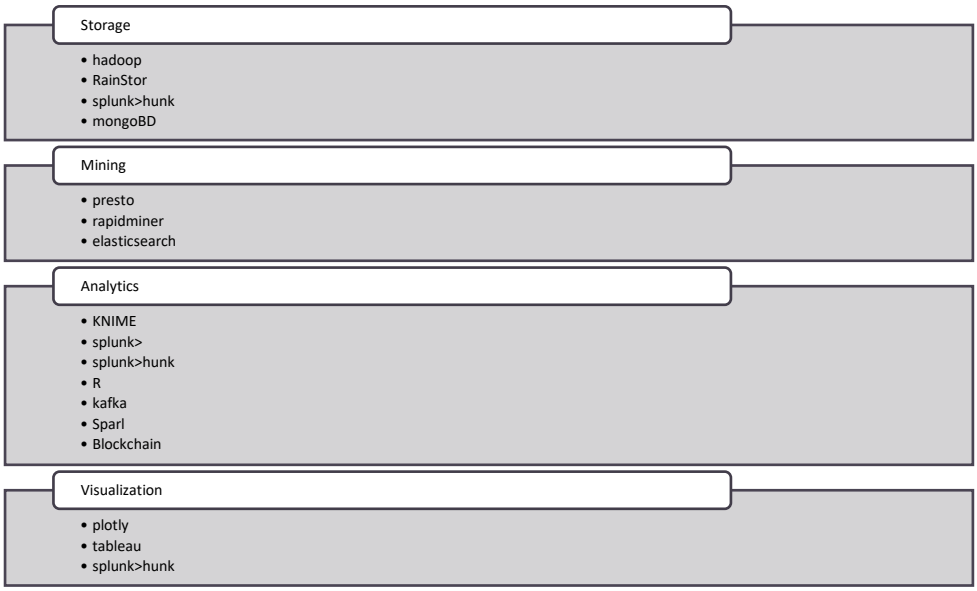


Figure 1 – List of technologies applied in Big Data. Available source: https://www.edureka.co/blog/top-big-data-technologies/

f. Big Data Analytics Lifecycle

Due to the characteristics (3 V´s) previously described of Big Data, Big Data analysis differs from traditional data analysis - requiring organizing the activities and tasks involved with acquiring, processing, analyzing and repurposing data. A specific proposal for data analytics lifecycle organizes and manages the tasks and activities associated with the analysis of Big Data. The Big Data analytics lifecycle can be divided into the following nine stages, as shown in figure 2 below:

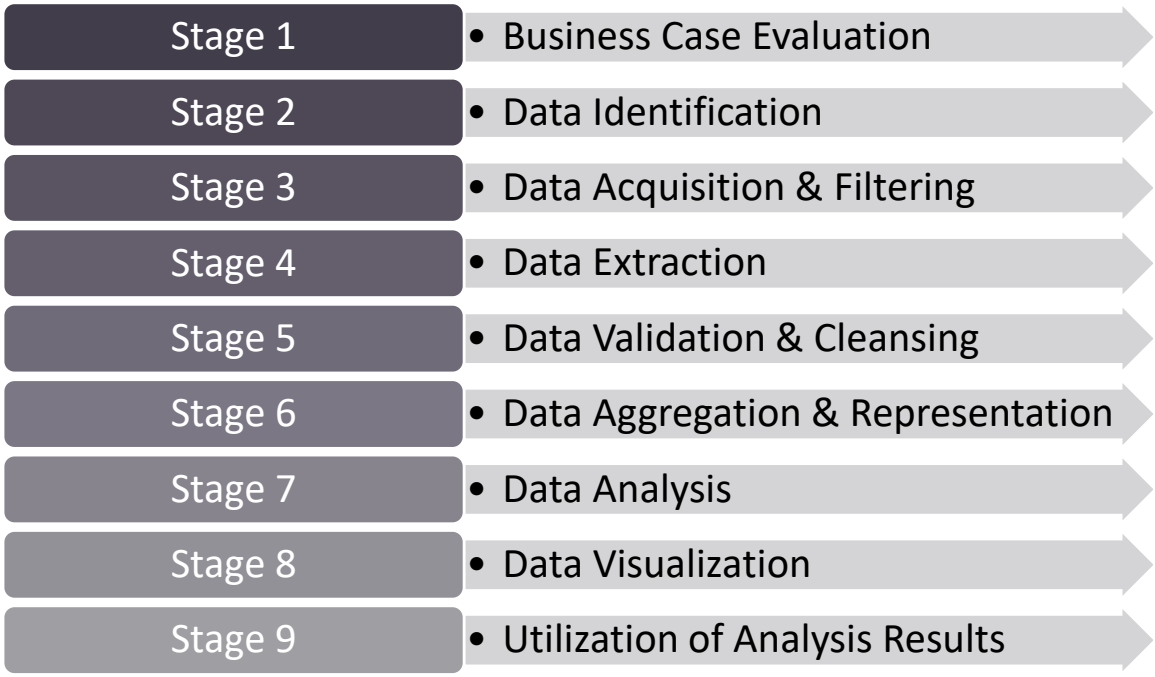


Figure 2 – Stages for the Big Data lifecycle. More information about each step can be found in the following link: https://www.informit.com/articles/article.aspx?p=2473128&seqNum=11.

2. Big Data Applications (Jiang et al., 2017; Moghadam & Colomo-Palacios, 2018; Ristevski & Chen, 2018)

There is a vast diversity of applications of big data analytics. In fact, throughout the process of gathering and analyzing data and supporting decision making with the best available evidence, they can:

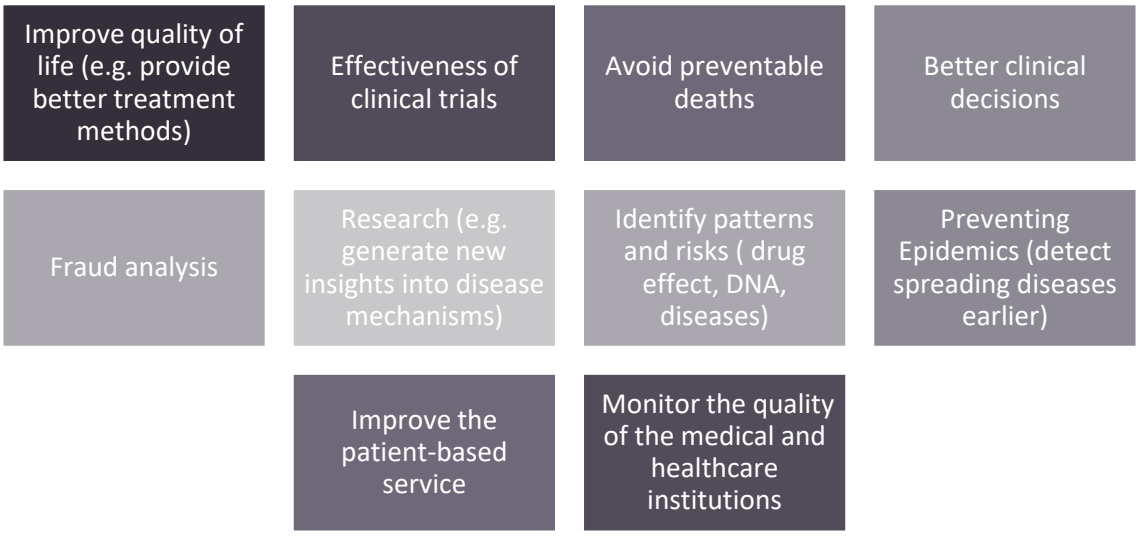


Figure 3 – Few applications of Big Data Analytics.

3. Big Data and secondary data (Benchimol et al., 2015; Cheng & Phillips, 2014)

Secondary data is defined as data used for different purposes of the ones they were originally collect for. Secondary data sources have been gaining attention as they provide an unparalleled retrospective opportunity to capture information across the entire system of health (see example 4).

There are several main sources of secondary data, including administrative databases, electronic health records, scientific literature, or scientific reports, surveys and internet-based data.

Some of the most visible advantages of secondary data are carrying out studies with less cost (most obvious one), in less time and with a larger sample size, which are also useful to raise new research hypotheses. Other advantages include the possibility of assessing long periods of time and large geographical areas (e.g. entire regions or nations). The analysis of large volumes of secondary data can overcome some limitations of “traditional” observational studies based on primary data, however concerns remain about these databases related to their possible heterogeneity and lack of

control during the collection process - with possible omission and lack of quality of the data. In fact, inherent to the nature of the secondary analysis, since any available data has been previously collected for other purpose its obvious this can lead to missing information on some important third variable or some specific segment of the population. Moreover, researchers who are analyzing the data are not usually the same individuals as those involved in the data collection process, so they may not be aware of some specifics that can lead to misinterpretation of the findings. Succinct documentation of important information about the validity of the data can partly mitigate this problem.

Data, data sources and types of data

We consider a clinical datum to be any single observation of a patient—e.g., a temperature reading, a red blood cell count, a past history of rubella, or a blood pressure reading. If a clinical datum is a single observation about a patient, clinical data are multiple observations (see example 3).

There is a broad range of data types/formats in the practice of medicine and the allied health sciences. They range from narrative, textual data to numerical measurements, genetic information, recorded signals, drawings, and even photographs or other images.

A database is a collection of individual observations (i.e., data) without any summarizing analysis. An Electronic medical record (EHR) system is thus primarily viewed as a database— the place where patient data are stored. When properly collected and analyzed with other data, these elements in the EHR provide information about the patient.

In fact, EHR has been described as a general term describing computer-based patient record systems. However, it is sometimes extended to include other functions like order entry for medications and tests, amongst other common functions which can start to actively support clinical care by providing a wide variety of information services. There are various types of EHR, namely:

- Electronic Medical Record (EMR) - includes all information (clinical and administrative) of one patient and focuses on relevant information for specific medical problem episodes;
- Electronic Patient Record (EPR) - is an organised collection of all records about an individual patient stored in the computer systems and databases of all the providers who have provided care to that patient within an enterprise;

- Virtual Patient Records (VPRs) - is a record that is not stored on any individual computer, but assembled dynamically, in real time, from various systems when needed;
- Electronic Health Record (EHR) - is a longitudinal record of patients' health. It combines information about patients' contacts with primary care and subsets of information associated with the outcomes of periodic care whether held in EMRs, EPRs or other information systems.
- Personal Health Record (PHR) - is a record that allows patient empowerment through personal management and sharing of personal health information.

Other examples of health data sources are provided in example 5.

Data from different sources are used for multiple purposes at different levels of the health care system. In fact, we can stratify data in individual level, health facility level data, population level data and public health surveillance. For example, the patient's profile, health care needs and treatment serves as the basis for clinical decision-making for individual clinical care while aggregated facility-level records from administrative sources enable health care managers to determine resource needs.

A data lake is a system or repository of data stored in its natural form (usually files). A data lake is typically a single repository of all business data, including raw copies of data from the source system and transformed data used for tasks such as reporting, visualization, analytics, and machine learning. Data Lakes increase agility and provide more opportunities for data exploration and proof of concept activities, as well as self-service business intelligence, within predefined privacy and security settings.

Remember:

DATA refers to raw numbers or other measures (objective facts about events) while **INFORMATION** refers to what emerges when data are processed, analyzed, interpreted, and presented. In other words, Information is Data transformed (contextualized, categorized, corrected, calculated, condensed) into a message. The key to any successful big data initiative is the ability to get information from the vast deluge of data, separating the noise from the signal.

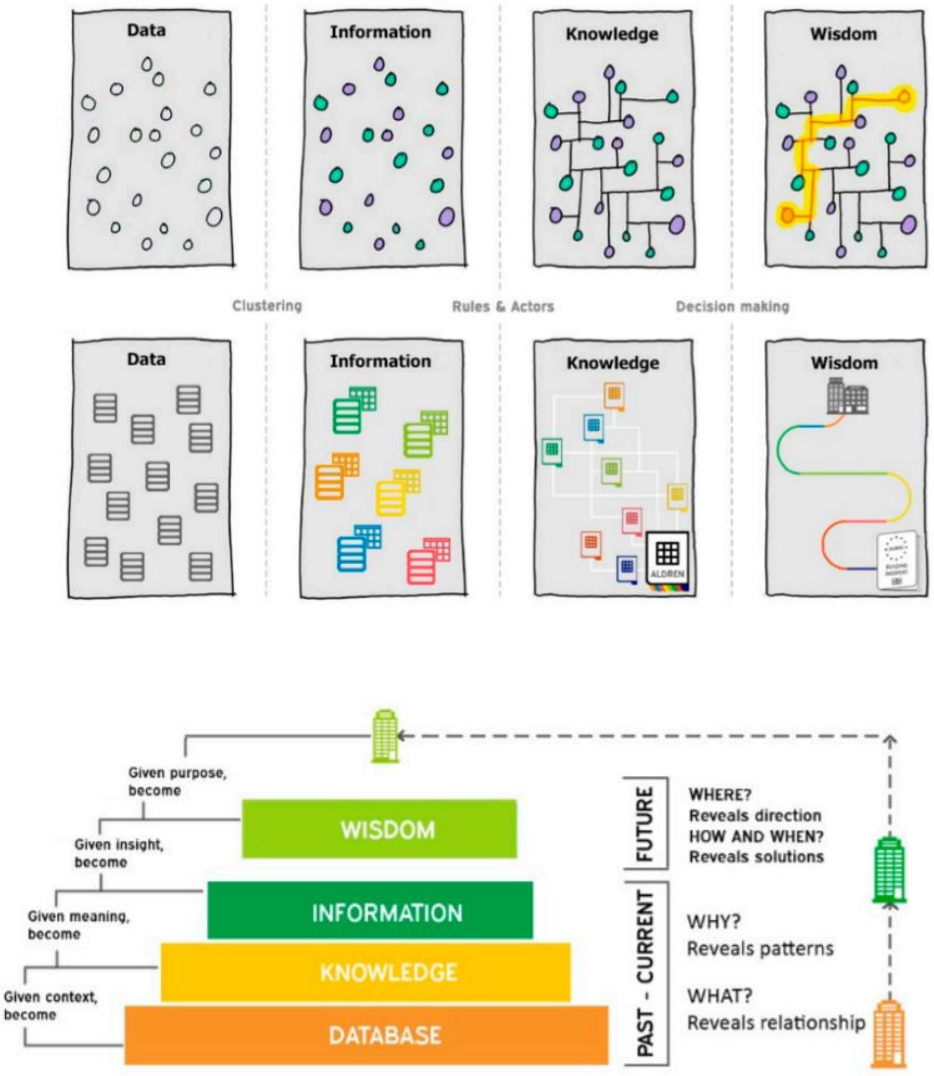


Figure 4 - Model for each level of information during data processing: from data do decision. Available source: (Sesana, Rivallain, & Salvalai, 2020)

4. Information systems in Health (WHO, 2008)

According to WHO, "the health information system collects data from the health sector and other relevant sectors, analyses the data and ensures their overall quality, relevance and timeliness, and converts data into information for health-related decision-making."

The health information system provides the underpinnings for decision-making and has four key functions:

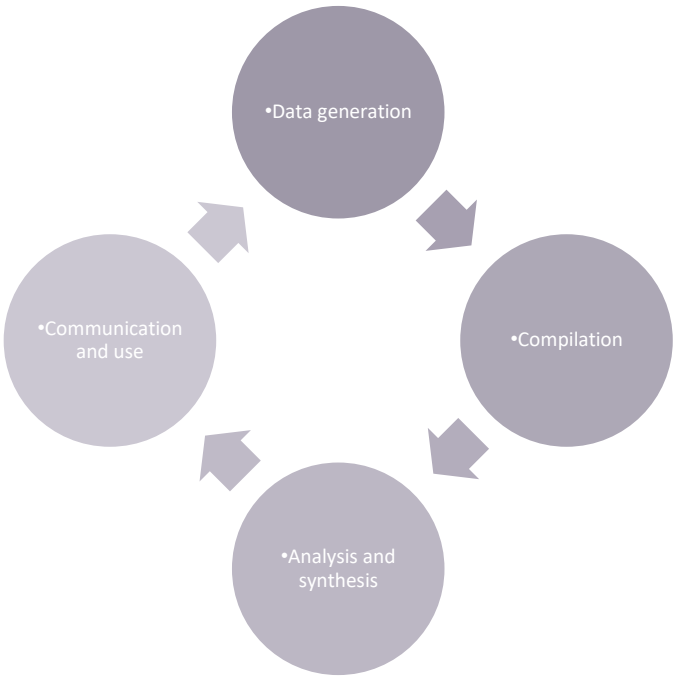


Figure 5 – The four key functions of health information system.

A robust information system infrastructure requires the ability not only to provide and make available quality data, but also to receive data, so they need to support bi-directional communication (of alerts, population health statistics and case or care management) - to inform clinicians and decision-makers in real-time. Information architecture (AI) refers to the logical configuration of various elements, including hardware, software, information flow and technical standards needed to support the information needs of users. Robust AI can increase the effectiveness and scope of its performance by integrating internal and external information systems. A fundamental component of information architecture is interoperability. Interoperability is defined as “the ability of a system to exchange electronic health information and use information from other systems without additional effort by the user”. Problems with data interoperability (ie, send, receive, find, and eventually be able to of use) restricts data exchange with other interested parties (see example 5).

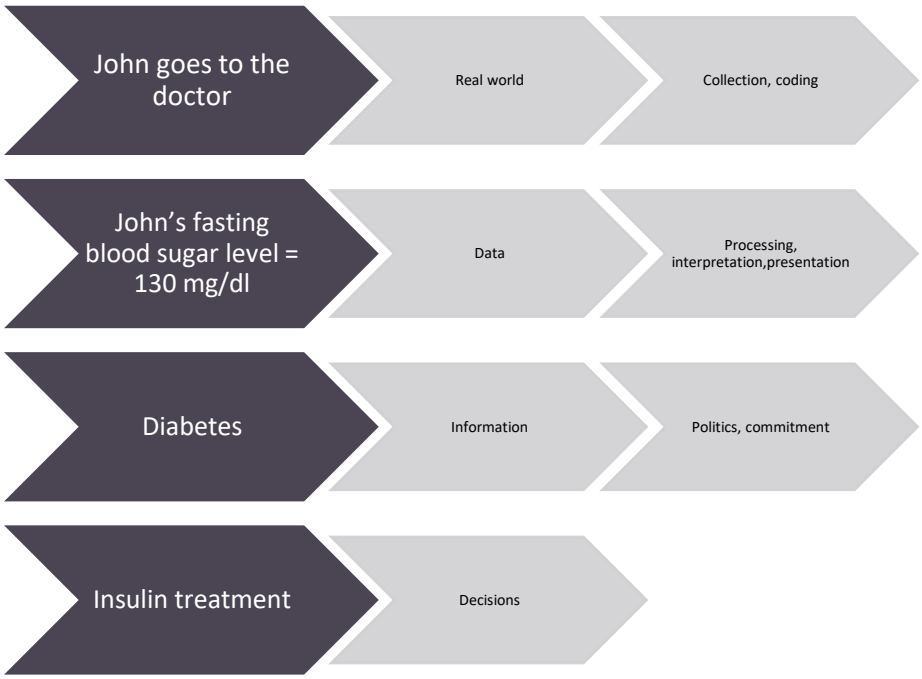
In fact, for an EHR to be effective, the data must be portable, which requires means to accurately and securely transfer the health data of a patient from one healthcare provider or facility to another

in a timely and efficient manner. Hence, the main focus of EHR technologies is standardization and connectivity. In fact, integration within and outside the healthcare facility constitutes a successful EHR deployment. Seamless connectivity between multiple, distributed systems in the healthcare continuum is the cornerstone of delivering a complete and accurate picture of the patient, their condition, treatment received, and subsequent outcomes. These connectivity challenges have been approached through the computerization of the world's healthcare operations and resulted in environments that are increasingly interoperable.

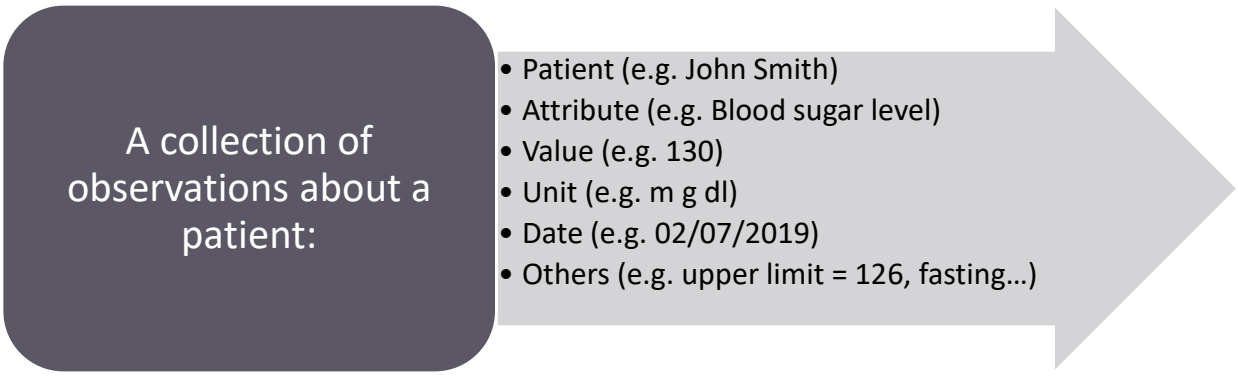
In conclusion, despite the introduction of EHR, which aim at recording and making accessible a patient's journey it is only recent advances in information technology that have created the infrastructure that allows these data to be used - by enabling data to be securely aggregated, stored, processed, and transmitted.

Examples and analogies

Example 1 – Understand the flow from real world to decisions



Example 2 – Applying the datum concept



Example 3 – Few types of data sources

- Electronic health record (EHR)
  - Patient's health records
  - Clinical provider controlled (not owned)
- Claims data
  - Administration
  - Billing
- Personal health record (PHR)
  - Wearables
  - Social
  - Patient controlled

Example 4 - Primary vs. Secondary use of Clinical Data

1. Clinical data is primarily used for clinical care
2. "Secondary" opportunities:
   a. Healthcare quality measurement
   b. Outcome comparison

c. Clinical research
d. Public health
e. Learning health system

Example 5 – EHR integration among multiple systems in the healthcare continuum

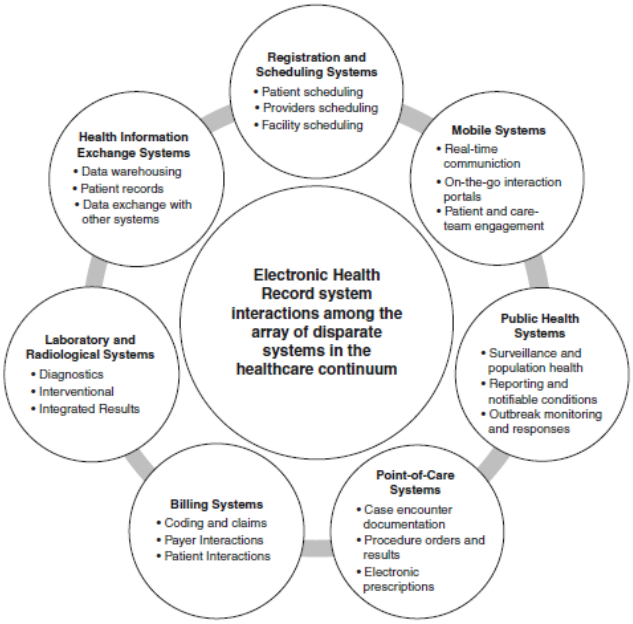


Figure 6 – EHR system interactions among multiple systems. Available at (Magnuson J.A., 2020)

Application and integration

Case Study 1 – Big Data Applications (Bhardwaj, Wodajo, Spano, Neal, & Coustasse, 2018; Wang & Hu, 2018)

As previously explained the ability to make timely and truly evidence-based informed decisions to provide more effective and personalized treatment while reducing the costs has been empowered by the introduction of big data analytics in healthcare. In fact, the ability to obtain and analyze big data can aid the identification of high-risk individuals, inform more effective treatments, and select cost reduction areas across the health care system. The application of big data analytics is vast within health care and goes beyond management of chronic diseases, management of resources to

management of acute public health situations (e.g., as foreseen with COVID-19). Many of these applications are listed in the following article. (Bhardwaj et al., 2018)

Nevertheless, acquiring such novel information is not free of limitations. In fact, it requires the integration of multiple kinds of datasets as well as information from many sources (questionnaire interviews, standard clinical tests, and modern sources such as electronic medical records, mobile apps, and wearable devices). As a case study from clinical perspective, the following paper provides an overview while highlighting some of these limitations, on how combined multiple data sources integration and applied big-data analytics may potentially inform personalized nutrition interventions prevention and management of type 2 diabetes.(Wang & Hu, 2018)

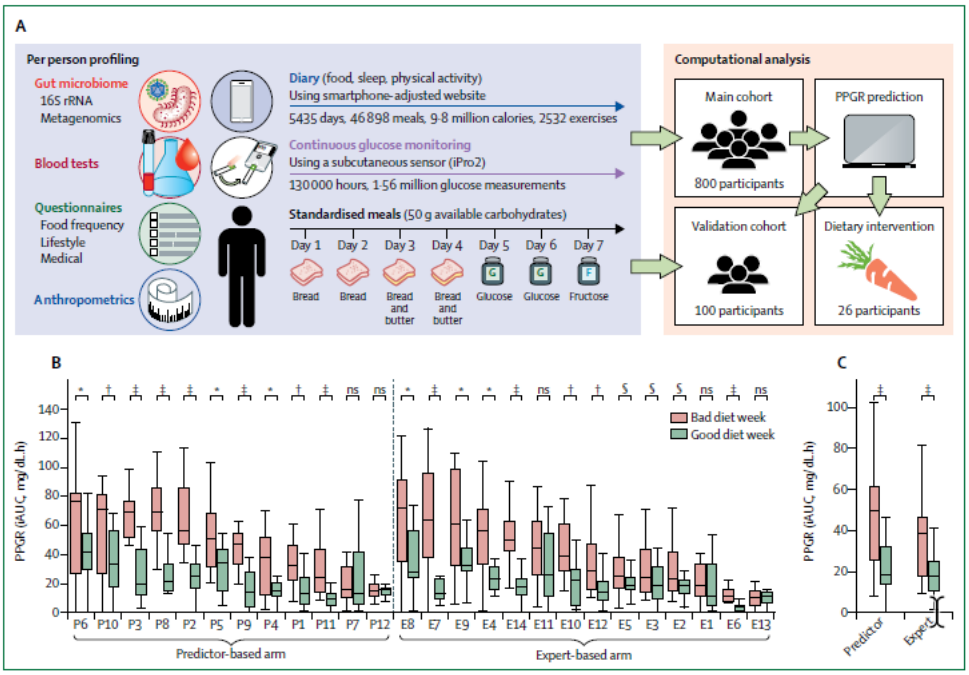


***Figure 2*: Personalised nutrition in reducing postprandial blood glucose**
(A) Illustration of the experimental design of Zeevi and colleagues.[7] (B) Left, mean postprandial glycaemic responses to personalised dietary intervention (good diet) *vs* control diet (bad diet), and right, traditional dietary advice (good diet) *vs* control diet (bad diet) at each timepoint of intervention period (by weeks). (C) Left, mean postprandial glycaemic responses to personalised dietary intervention *vs* control diet, and right, traditional dietary advice *vs* control diet. PPGR=postprandial glucose response. Good diet=meals predicted to have low postprandial glycaemic responses. Bad diet=meals predicted to have high postprandial glycaemic responses. *p<0·001, †p<0·01, ‡p<0·05, §p<0·1, ns=not significant (Mann-Whitney *U* test). iAUC=incremental area under the curve. Adapted from Zeevi and colleagues,[7] by permission of Elsevier.

Figure 7 – Design for the case study 1: multiple sources of information. Available at (Wang & Hu, 2018)

References for further information and areas on inquiries

Magnuson J.A. DBE. Public Health Informatics and Information Systems: Springer International Publishing; 2020.

Shortliffe E, Cimino J. Biomedical Informatics: Computer Applications in Health Care and Biomedicine 2014.

Cruz-Correia R, Rodrigues P, Freitas A, Almeida F, Chen R, Costa-Pereira A. Data Quality and Integration Issues in Electronic Health Records. 2009. p. 55-95.

Lesson plan 2: Big Data Analytics

Foundational knowledge

1- Introduction to Descriptive, Predictive and Prescriptive analysis

Science and technology are developing at an unparalleled rate. The convergence of new treatments, diagnostics, wearables, sensors and connectivity is generating enormous amounts of data. As the amount of data available increased dramatically so the motivation for different analytical approaches. In fact, more data available does not always mean more knowledge to be used in decisions, as we need automatic methods to exploit such data.

Analytical approaches can be defined three categories, namely descriptive, predictive, and prescriptive:

• Descriptive analytics refers to summarization of datasets making them interpretable to researchers.

•Predictive analytics approaches provide estimates on the likelihood of future events or outcomes - or are used to address gaps or missing information.

• Prescriptive analytics goes beyond descriptive and predictive analytics by attempting to quantify the impact of future decisions before these decisions are made.

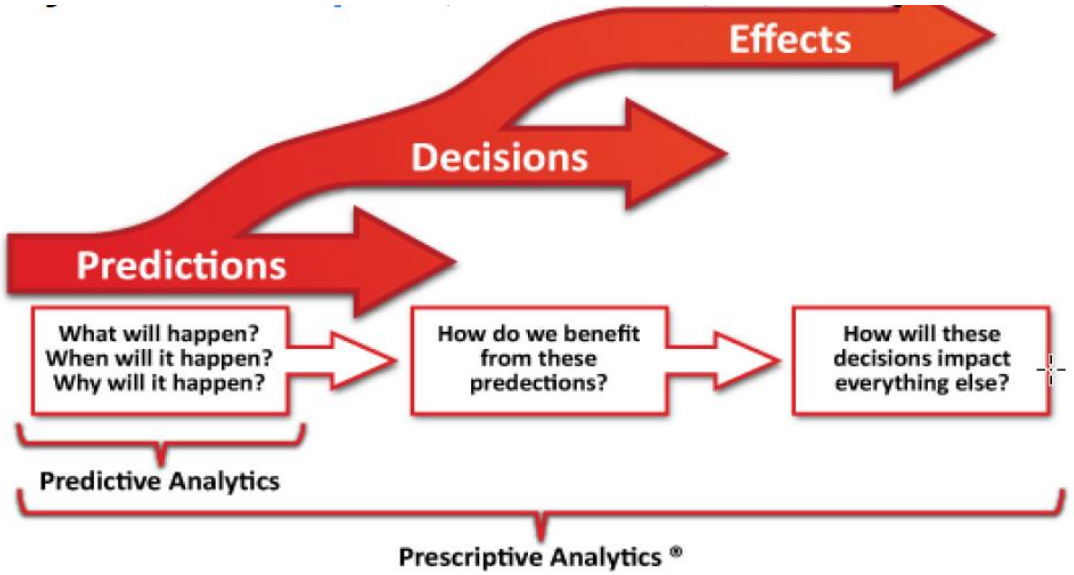


Figure 8 – Analytical approaches to process data. Image source available at:https://en.wikipedia.org/wiki/Prescriptive_analytics

2 - Computational methods for large databases (analytical and modeling techniques) (Magnuson J.A., 2020)

Computational power and approaches based on bioinformatics tools and algorithms, machine learning (ML) or artificial intelligence (AI) are gaining access to the health care systems. Likewise, our possibilities for evidence generation are growing.

ML techniques can be categorized into supervised and unsupervised learning approaches.

Machine Learning Main Techniques:

- Supervised Techniques
  - Classification
  - Regression
- Unsupervised Techniques
  - Clustering
  - Association Rules

Supervised and Unsupervised Learning Algorithms

The availability of the outcome of interest is the big difference between both approaches. In fact, while in supervised learning algorithms the outcome is given to the algorithm – which can use it as gold standard to convert the input features into the outcome – in unsupervised learning approaches are methods where an algorithm must learn to model the underlying distribution of data elements given input features, but no outcome variable. Supervised methods can be costly and resource intense as they may require human expert input for defining and preparing a gold standard (i.e. the output label). In contrast, unsupervised methods rely purely on the quantity and quality of data for the training process - do not require the manual cost and effort required to develop a gold standard which can lead to weaker performance.

For supervised approaches Classification (which can predict a discrete or categorical output variable) and Regression (predicts numerical continuous output variable) models are the major categories. Some classification models are (1) Simple logistic – uses a logistic function which is used to predict the outcome variable; (2) Support vector machines - identifies an optimal hyperplane (a subspace whose dimension is -1 of its ambient space) capable of separating data into each outcome; (3) Decision trees – generally predicts the value of an outcome by learning decision rules inferred from the training dataset. Among examples of Regression Algorithms are simple logistic regression and random forest regression.

Overall, the process in supervised ML encompasses the following steps:

1. During training the model is given both the features and the labels and learns how to map the former to the latter.
2. A trained model is evaluated on a testing set, where we only give it the features and it makes predictions.
3. Then, the predictions are compared with the known labels for the testing set to calculate accuracy.

K-means clustering and Hierarchical clustering are the most widely known unsupervised learning algorithms. The first approach seeks to group each observation into a subset of clusters where each observation belongs to the cluster with the nearest mean value. In contrast, the second uses an approach which seeks to build out a hierarchy of clusters, which can be agglomerative (each

individual instance starts as a separate cluster, with pairs of clusters merging as instances traverse up the hierarchy) or divisive (all observations start with one cluster and splits are performed as instances traverse down the hierarchy). Please see example 6.

3 - Artificial intelligence in healthcare: Fundamentals

During the process of designing and selecting the best model approach, there are few methods or approaches, namely: train and test, cross -validation and train, validation and test. In the first method a dataset is randomly split into two sets, a larger training dataset used to train a decision model, and a smaller test dataset used to test the newly trained model. In the Cross-validation approach the dataset is split into k many randomly selected subsets (using a pre-defined size *k*). In each stage, one subset plays the role of validation set whereas the other remaining parts (K-1) are the training set. For each stage, it involves removing part of the data, then holding it out, fitting the model to the remaining part, and then applying the fitted model to the data that we've held out. Performance results for each iteration are averaged to produce less variable performance results. In the last approach, dataset is randomly split into train, validation, and test sets. The training dataset is used to train the decision model. The validation dataset is then used to iteratively test the decision model, and update its parameters for optimal performance. Once model parameters have been configured for optimal results, the model is then evaluated using the test dataset (see figure 9).

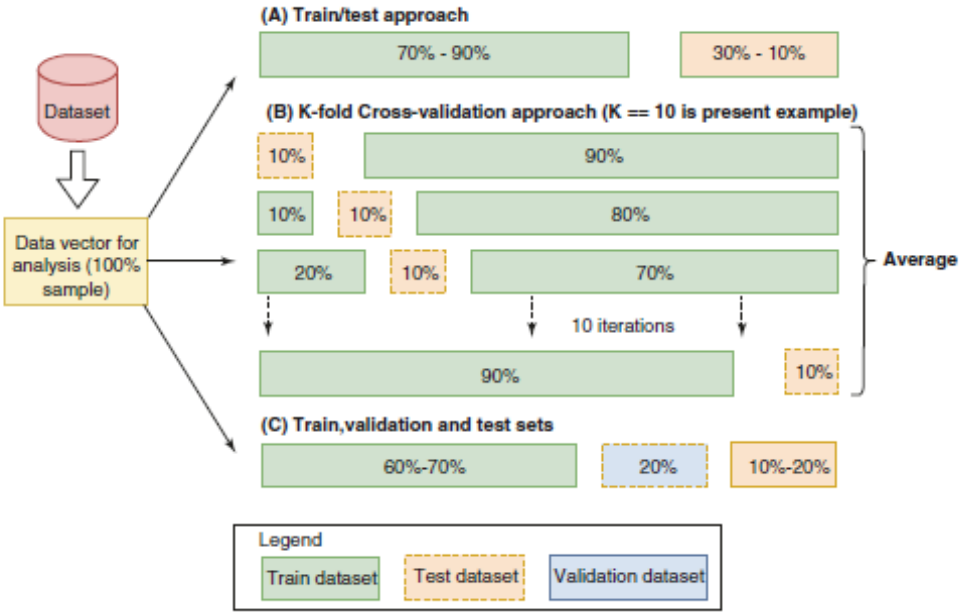


Figure 9 - Methods to design and develop an algorithm of machine learning. Image available at (Magnuson J.A., 2020)

Likewise, it is essential to evaluate the performance characteristics of a decision model. For this some used performance metrics are: sensitivity (i.e. the proportion of actual positives that are correctly identified); specificity (i.e. the proportion of actual negatives that are correctly identified); precision (i.e. the proportion of positive identifications that are correct); F1-score (i.e. accuracy measure representing an average used for numbers that represent rate or ratio between precision and sensitivity; and Area under the Receiver Operator Characteristic curve (AUC ROC) which demonstrates through a graphical plot the diagnostic performance of a classification model across various threshold configurations (this score can range between 0-1).

To wrap up, a model learns relationships between the inputs (features) and outputs (labels) from a training dataset. Models can take many shapes such as logistic regression – described in the previous section.

However, ML algorithms are not free of constraints. While building these models, researchers should be aware and concern on some challenges. Overfitting, broadly speaking and in contrast to underfitting, means the training fits exactly against its training data resulting in inability to generalize to unseen datasets. This happens when noise (irrelevant or incorrect data elements) is included in the dataset. In other hand, underfitting means the model has not captured the underlying logic of the data, thus it is unable to properly perform across both the current and new datasets. Finally,

class imbalance occurs when classes are not present in proportion across the dataset, i.e., the occurrence of one of the classes is very high compared to the other classes present (there is a bias or skewness towards the majority class present in the target). Two sampling methods are usually used to address this problem: oversampling (supplementing with observations of minority instances) and undersampling (removing instances of the majority class).

4 - Artificial intelligence in healthcare: Issues

In a recent report on Ethics Guidelines for Trustworthy Artificial Intelligence (AI), AI systems were defined as " software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." Likewise, according to the same guidelines a "trustworthy AI has three components, which should be met throughout the system's entire life cycle; it should be:

(1) lawful, complying with all applicable laws and regulations;

(2) ethical, ensuring adherence to ethical principles and values, and

(3) robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm."

There are still many open ethical, scientific and technological challenges to build the capabilities that would be needed to achieve a trustworthy AI, especially if, for example, we consider a general AI system which is intended to be a system that can perform most activities that humans can do, such as common-sense reasoning, self-awareness, and the ability of the machine to define its own

purpose. Nevertheless, currently deployed AI systems are examples of narrow AI (systems that can perform only one or few specific tasks).

Moreover, ensuring a trustworthy AI requires efforts during its whole life cycle as such systems can inherited many issued. Constrains may refer to data bias and/or model explicability. In fact, since AI systems rely on data to perform well, if the training data is imbalanced or biased the model will not have the capacity to generalize. Explicability refer to the capacity to provide a form of explanation for the systems decisions, i.e, transparency in terms of understanding how they make decisions.

To sum up, overall, achieving a Trustworthy AI must be translated into concrete requirements (i) Respect for human autonomy, (ii) Prevention of harm, (iii) Fairness and (iv) Explicability. These requirements are applicable to different stakeholders intervening on AI systems' life cycle - developers, deployers and end-users, as well as the broader society and are listed below in figure 10.

Figure 10 – Requirements to implement and evaluate throughout AI system´s lifecycle. Available at: (Comission, 2019)

Examples and analogies

Example 6 – In the following practical course you will be able to understand how such concepts apply:

https://www.kaggle.com/learn/intro-to-machine-learning

Application and integration

For a hands-on practical code experience you can visit the following open crash course:

https://machinelearningmastery.com/start-here/

References for further information and areas on inquiries

Comission E. Ethics guidelines for trustworthy AI. 2019.

Magnuson J.A. DBE. Public Health Informatics and Information Systems: Springer International Publishing; 2020.

Lesson plan 3: Data driven decision making

Foundational knowledge

1- Principles and definition of Data Governance

There is still no consensus for a proper definition of Data Governance.

IBM defined data governance "as a discipline of quality control to add new rigor and discipline to the process of managing, using, improving and protecting organizational information". (IBM, Data Governance) Recently, de EU splits this concept and defines firstly "data governance that entails defining, implementing and monitoring strategies, policies and shared decision-making over the management and use of data assets" and secondly "data policies are a set of broad, high level principles which form the guiding framework in which data assets can be managed." More specifically, both concepts "aim to provide guidance, assurance and support to transform the data-driven organization by

defining clear roles and responsibilities; and introducing common principles, guidance and working practices that provide the foundation for harmonized and coordinated data management across the organization."

To this end, it is defined 3 organizational hierarchical levels, namely:

1. Operational
    a. Implementation of data policies and accountable for local decisions about data
    b. Whenever necessary, issues are escalated to the managerial level for resolution.
2. Managerial
    a. Responsible for developing and implementing data policies at corporate level and local level.
    b. It monitors progress, reports to the strategic level and refers to them any issues and matters that are beyond its decision-making power or mandate.
3. Strategic
    a. Defines the long-term vision, gives direction, oversees progress, takes strategic decisions, and acts as the highest point of reference for issues and matters related to data governance and data policies.

Overall, data governance enables a greater degree of transparency, auditability, and accountability of the organization's data assets. Therefore, a strategy for a data governance program should cover not only the organizational, data management priorities, but also aspects as legislation and regulatory compliance and data culture and structures within the organization.

The principles for implementing data governance strategy are listed below:

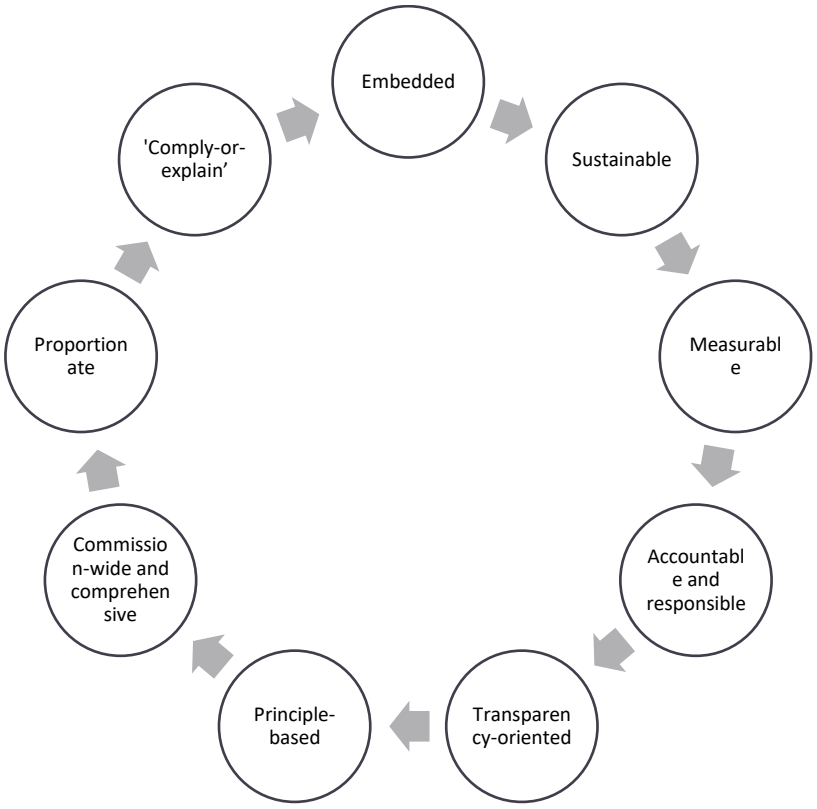


Figure 10 – Principles for implementing a data governance strategy.

2 - Challenges of Data Governance (Comission, 2020)

(1) Data Quality

Researchers, decision makers and data scientists must consider possible biases and limitations, regarding the accuracy and integrity of secondary data, where the 6 dimensions of data quality stand out:

1- Accuracy measures the degree in the which in what is measured represents a relationship with the real world;

2- Completeness represents if all necessary fields are registered;

3- Consistency is related to the frequency of filling in the various data fields;

4- Temporality measures the time difference in which the use of data is expected and in which they can effectively be used;

5- Uniqueness refers to the duplication or contradiction of records;

6- Validity is linked to the correct format, data type and consistency with pre-defined parameters.

Best practices on data quality management should focus on data quality planning, control and monitoring, assurance, and improvement (figure 11).

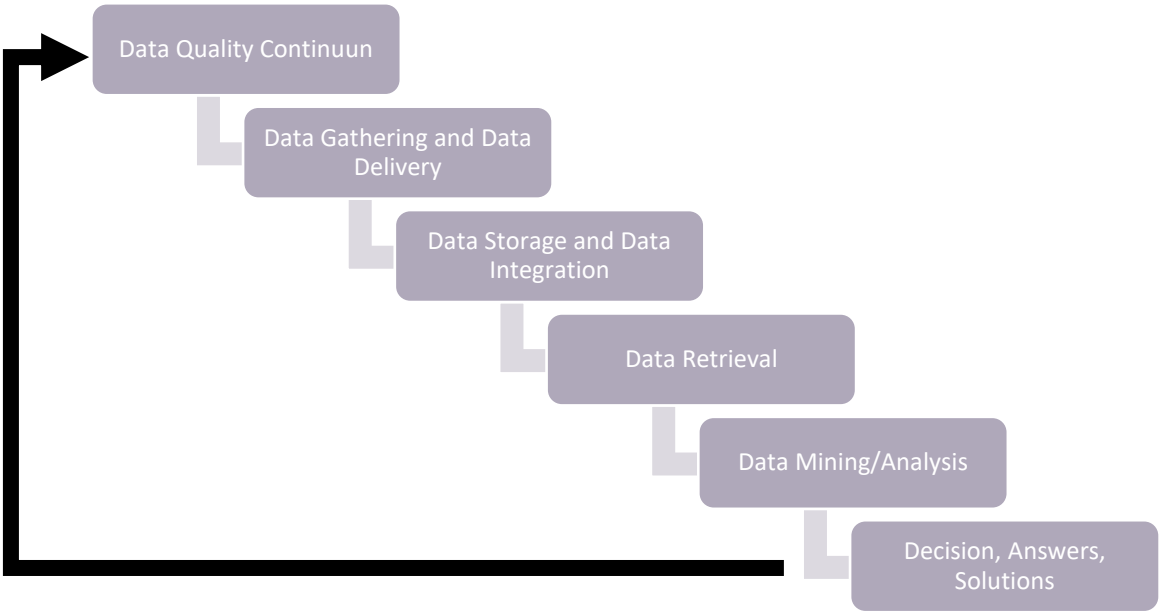


Figure 11 – Flow for data quality management.

(2) Protection and Information Security

Privacy and Confidentiality

With big data comes big risks and challenges, among them significant questions about patient privacy and confidentiality. To this end, the European Commission (EC) proposed a key reform of the EU legal framework, which led to the draft of a new European regulation on the protection of personal data.

European Data Protection Regulation (GDPR) and Health Data

In 2018, EU Regulation 2016/679 of the European Parliament of the council of 2016 states to harmonize data privacy laws across Europe (Commission, 2016). Accordingly 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" (Commission, 2016).

The GDPR creates a distinction between personal data and special categories of personal data, which merit higher protection, such as personal health data which includes genetic data, biometric data or any data concerning health.

Personal data concerning health includes "all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person (…); a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test" (Commission, 2016). In addition, it establishes the general principle of the prohibition of processing health data. The GDPR broadly defines processing as including operations on personal data as well as data collection, storage, use, disclosure, and destruction. The GDPR rules are rooted in six data processing principles defined in GDPR Article 5, which must be followed:

1. Lawfulness, fairness, and transparency— requiring lawful, fair, and transparent data processing
2. Purpose limitation—requiring data be processed consistently with the purpose for which it was collected
3. Data minimization—limiting processing to what is necessary for a given purpose
4. Accuracy-requiring "every reasonable step" to ensure that data are accurate

5. Storage limitation—limiting the storage of identifiable data
6. Integrity and confidentiality—requiring appropriate security for processing personal data

Privacy and confidentiality

Privacy and confidentiality of protected health information (information that can be used to identify an individual) are major drivers to system and data security. Privacy is viewed from the perspective of the individual, it reflects the right to keep one's information private (undisclosed). Confidentiality is viewed from the perspective of those entrusted with that information, and their duty to keep the information private (see example 7). Confidentiality - ensuring confidentiality begins with a data classification program and requires the highest levels of protection - if disclosed, disrupted, or stolen, would cause considerable damage to an organization. One common and effective method of ensuring confidentiality is to encrypt the data.

(3) Data Interoperability and Standards

According to CDC (CDC, 2021), effective interoperability of healthcare data ensures that electronic health information is shared appropriately between healthcare and public health partners in the right format, through the right channel at the right time. Data Interoperability's goal is to improve communication between entities and ultimately improve patient care. In fact, some benefits include:

- Bi-directional communication between state public health departments and clinical care providers and between organizations
- Standardized data elements for data exchange
- Improved efficiency across the healthcare and public health system

But…Health data is inconsistent, unstructured, and complex, raising difficulties in applying interoperability principles!

Currently among the available technical specifications for achieving interoperability, HL7 is the most widely used syntactic standard in healthcare and in recent years HL7® has developed a new product called Fast Healthcare Interoperability Resources (FHIR) - the latest development in standards for health information exchange developed by HL7.

A key strength of FHIR is that it uses an application programming interface (API) based approach to facilitate exchange as well as integration of clinical data thus facilitating interoperability.

Nevertheless, achieving interoperability still faces many challenges, highlighting funding, the sparce variety of platforms and HIE, complex legal and regulatory environment and workforce.

(4) Data Governance and Management

While governing a data program, a holistic approach should set the basis for best practices during such process. Applying core principles which govern the life cycle of information (as previously discussed) may set the basis for an efficient data management strategy in the health ecosystem.

Examples and analogies

Example 7 – Example on applying the concept of privacy and confidentiality:

"For example, a Public Health (PH) professional engaged in Sexually Transmitted Disease (STD) contact tracing might be working to identify and help the sexual partner(s) of an HIV/AIDS individual. The PH professional will inform the contact(s) that they may have been exposed to an STD, and recommend testing/treatment. But if the original individual wishes to keep their information private, then the partner(s) will not be told who referred them—that information will be kept confidential."

Application and integration

Case Study 2 (Ozaydin, Zengul, Oner, & Feldman, 2020)

As previously stated data Governance requires multiple fields of intervention. Application of such dimensions range from projects applied to address integration, cleansing, interpretation, and aggregation of raw data from multiple data sources.

In this section we will use the article to provide an overview of health life cycle.

Current uncoordinated and isolated efforts on such disparate datasets can be wasteful due to inability to reproduce findings. Information technology (IT) infrastructure is crucial to unveil the

potential of analytics and address data governance constraints, namely in terms of interoperability. This is especially important when multiple databases are available and a proper design of infrastructure architecture may improve data delivery to health services researchers – thus, making it possible for collaboration on innovative and rigorous research. Data warehouses usually serve as the infrastructure for running institutional data analytics and business intelligence (BI) systems.

In this case study, it is provided theoretical underpinnings of the processes and methodologies in developing a data warehouse system as an infrastructure to support health services research.

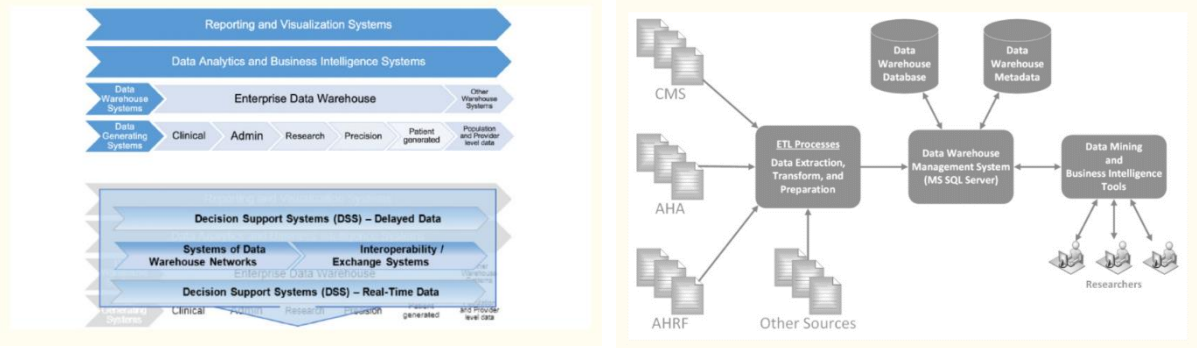


Figure 12 – Processes and methodologies for developing a data warehouse system as an infrastructure to support health services research. Images extracted from paper of Case Study 2.

It addresses existing constrains in data quality (e.g., inefficiencies, disparate and unnecessary duplication of efforts, and the lack of harmony among health services researchers) during all health data lifecycle. To this end, it is presented and discussed a design process (the application of 4 phases of a conceptual iterative process model) for the implementation of HRADIS - a full-service data warehouse integrating frequently used health services research data sources, processes, and methods along with a variety of data analytics and visualization tools.

Other example that has been gaining attention from the scientific community is the FAIR4Health. For more information visit https://www.fair4health.eu/.

References for further information and areas on inquiries

Comission E. Data governance and data policies at the European Commission. 2020.

Commission, E. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

4. Appendices

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., . . . Langan, S. M. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med, 12*(10), e1001885. doi:10.1371/journal.pmed.1001885

Bhardwaj, N., Wodajo, B., Spano, A., Neal, S., & Coustasse, A. (2018). The Impact of Big Data on Chronic Disease Management. *Health Care Manag (Frederick), 37*(1), 90-98. doi:10.1097/hcm.0000000000000194

CDC. (2021). Goals and Benefits of Data Interoperability. Retrieved from https://www.cdc.gov/datainteroperability/goals-and-benefit.html

Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: opportunities and implementation. *Shanghai archives of psychiatry, 26*(6), 371-375. doi:10.11919/j.issn.1002-0829.214171

Comission, E. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Comission, E. (2020). *Data governance and data policies at the European Commission*. Retrieved from https://ec.europa.eu/info/publications/data-governance-and-data-policies-european-commission_en

Commission, E. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Cruz-Correia, R., Rodrigues, P., Freitas, A., Almeida, F., Chen, R., & Costa-Pereira, A. (2009). Data Quality and Integration Issues in Electronic Health Records. In (pp. 55-95).

Davenport, T., & Harris, J. (2007). *Competing on Analytics: The New Science of Winning*.

Gao, X., & Yu, J. (2020). Public governance mechanism in the prevention and control of the COVID-19: information, decision-making and execution. *Journal of Chinese Governance, 5*, 178 - 197.

Halevi, G. (2014). Research Assessment: Review of methodologies and approaches. *Research Trends*.

Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., & Glasziou, P. P. (2014). *Decision Making in Health and Medicine: Integrating Evidence and Values* (2 ed.). Cambridge: Cambridge University Press.

IBM. (Data Governance). Retrieved from https://www.ibm.com/analytics/data-governance

Janssen, M., & van der Voort, H. (2020). Agile and adaptive governance in crisis response: Lessons from the COVID-19 pandemic. *International Journal of Information Management, 55*, 102180. doi:https://doi.org/10.1016/j.ijinfomgt.2020.102180

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., . . . Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology, 2*(4), 230-243. doi:10.1136/svn-2017-000101

Magnuson J.A., D. B. E. (2020). *Public Health Informatics and Information Systems*: Springer International Publishing.

Maissenhaelter, B. E., Woolmore, A. L., & Schlag, P. M. (2018). Real-world evidence research based on big data. *Der Onkologe, 24*(2), 91-98. doi:10.1007/s00761-018-0358-3

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *Int J Med Inform, 114*, 57-65. doi:10.1016/j.ijmedinf.2018.03.013

Moghadam, R. S., & Colomo-Palacios, R. (2018). Information security governance in big data environments: A systematic mapping. *Procedia Computer Science, 138*, 401-408. doi:https://doi.org/10.1016/j.procs.2018.10.057

Ozaydin, B., Zengul, F., Oner, N., & Feldman, S. S. (2020). Healthcare Research and Analytics Data Infrastructure Solution: A Data Warehouse for Health Services Research. *J Med Internet Res, 22*(6), e18579. doi:10.2196/18579

Pastorino, R., De Vito, C., Migliara, G., Glocker, K., Binenbaum, I., Ricciardi, W., & Boccia, S. (2019). Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *Eur J Public Health, 29*(Supplement_3), 23-27. doi:10.1093/eurpub/ckz168

Ristevski, B., & Chen, M. (2018). Big Data Analytics in Medicine and Healthcare. *Journal of integrative bioinformatics, 15*(3), 20170030. doi:10.1515/jib-2017-0030

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed.), 312*(7023), 71-72. doi:10.1136/bmj.312.7023.71

Schlomer, B. J., & Copp, H. L. (2014). Secondary data analysis of large data sets in urology: successes and errors to avoid. *J Urol, 191*(3), 587-596. doi:10.1016/j.juro.2013.09.091

Sesana, Rivallain, & Salvalai. (2020). Overview of the Available Knowledge for the Data Model Definition of a Building Renovation Passport for Non-Residential Buildings: The ALDREN Project Experience. *Sustainability, 12*, 642. doi:10.3390/su12020642

Shabani, M. (2021). The Data Governance Act and the EU's move towards facilitating data sharing. *Molecular systems biology, 17*(3), e10229-e10229. doi:10.15252/msb.202110229

Shortliffe, E., & Cimino, J. (2014). *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*.

Wang, D. D., & Hu, F. B. (2018). Precision nutrition for prevention and management of type 2 diabetes. *The Lancet Diabetes & Endocrinology, 6*(5), 416-426. doi:https://doi.org/10.1016/S2213-8587(18)30037-8

Watson, H. (2019). Update Tutorial: Big Data Analytics: Concepts, Technology, and Applications. *Communications of the Association for Information Systems, 44*, 364-379. doi:10.17705/1CAIS.04421

WHO. (2008). HEALTH INFORMATION SYSTEMS Retrieved from https://www.who.int/healthinfo/statistics/toolkit_hss/EN_PDF_Toolkit_HSS_InformationSystems.pdf