Data Analytics

Subgroup report

Paolo Alcini - EMA (subgroup lead)
Gianmario Candore – EMA (subgroup lead)
Marek Lehmann - EMA
Luis Pinheiro - EMA
Antti Hyvärinen - Fimea
Hans Ovelgonne – CBG-MED
Mateja Sajovic - JAZMP
Panagiotis Telonis - EMA
Kevin Horan – HPRA (until May 2018)
Massimiliano Falcinelli – EMA (from September 2018)

See websites for contact details
Heads of Medicines Agencies www.hma.eu
European Medicines Agency www.ema.europa.eu

An agency of the
European Union

Table of content

Table of Figures

Executive summary

It is tempting to assume that the ever-increasing availability of electronic health data will transform the science of medicine and the way healthcare is provided. It is, however, essential to remember that *data by itself does not provide value*: it needs to be *analysed*, interpreted and acted upon.

This report is a deliverable of the HMA EMA Big Data task force 'Data analytics' subgroup, and focuses on four activities (*standardisation, information technology, data manipulation* and *artificial intelligence*) that are interrelated and necessary to extract insights and knowledge from data.

Each of the four topics offers possibilities, from different perspectives, to engage with the challenges posed by big data and to create capability and capacity to guide and profit from the opportunities. This, in turn, will improve regulation either by being *more efficient* (freeing resources for core regulatory activities) or by *enhancing support for regulatory science*[1] *and decision making*.

The main aim of the 'Data analytics' subgroup is to be an *enabler*, from a *technological and methodological* point of view, for the other subgroups of the task force by both facilitating a better understanding of the methodologies and techniques which the regulatory network should focus on, and also providing a practical list of recommendations. A greater understanding will avoid perceiving the techniques as a black box and also facilitate in asking the questions necessary to comprehend whether they are appropriate for the data used and to provide a clear path for validation when necessary.

The report focuses on the main concepts, assuming the reader is a scientist or assessor looking for guidance, or a trained analyst seeking clarifications and a structured presentation of the main elements to consider when analysing big data[2].

Data Standardisation

Every Standard Development Organisation (SDO) has its own definition of standards. For ISO, for instance: *"An international standard provides rules, guidelines or characteristics for activities or for their results, aimed at achieving the optimum degree of order in a given context. It can take many forms. Apart from product standards, other examples include: test methods, codes of practice, guideline standards and management systems standards".*

Several SDOs in regulatory medicines are also members of the Joint Initiative council (JIC). Some of its members with which the EU regulatory network has interacted in the past are: ISO, CEN, HL7, CDISC, SNOMED International and GS1. In EU, the main standardisation framework is legalised by the European Parliament and Council Regulation (EU) No 1025/2012. Also, ICH has a key role.

In the context of data there are standards for defining: data elements, terminologies, unit of measurements, data models, ontologies, data acquisition/collection processes, file formats, electronic messaging and processes related to data analytics.

The EU Regulatory Network has a level of heterogeneity that makes difficult to achieve a common interpretation of the data; in addition, there are challenges in sharing, integrating, quality assure and

[1] "Regulatory science is defined as the range of scientific disciplines that are applied to the quality, safety and efficacy assessment of medicinal products and that inform regulatory decision making throughout the lifecycle of a medicine. It encompasses basic and applied biomedical and social sciences, and contributes to the development of regulatory standards and tools" EMA Regulatory Science to 2025 https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf

[2] Technical details have not been reported in the main text, but the interested reader can find some of them described in the Annex A and Annex B.

analysing big data within the network. Standards can help to overcome these key issues helping to reduce costs by anticipating technical requirements and increase productivity. Standards can also augment data sharing by enabling the integration of silo-databases, as well as improve data quality, processes and foster good analytics. Finally, standards can ensure that data is collected and provided in a usable and coherent format.

Therefore, to benefit from these opportunities, some key foundational recommendations are:

To engage with SDOs and other organisations that are developing open standards (e.g. GA4GH).

To ensure that legislators incorporate, where possible, agreed standards when developing guidelines and regulations.

To adopt (and adapt when needed) international standards to foster harmonisation and improve efficiency, if the return of investment presents a positive benefit.

The European Network Data Board (EUNDB), consulting with other scientific and IT committees and working parties, should remain the forum for the EU regulators to raise and discuss standardisation needs and to develop consensus positions on the implementation of new standards fora (e.g. ISO plenary meetings).

If not already in place, national regulatory agencies should initiate direct partnerships with their national ISO standardisation bodies and EMA should develop similar partnerships with European wide bodies such as CEN.

Finally, some more specific recommendations are:

To foster the use of SPOR in clinical trial to univocally identify the substance under investigation.

To create a new standard in ISO or HL7 for creating an overarching application programming interface (API) for exchanging genomic data.

To explore if an ISO project could be undertaken to merge all the various initiatives in the bio-analytics omics domain and create a set of ISO standards in collaboration with CDISC and HL7.

To develop an API to retrieve/download data from the various social media platforms.

To contribute to the work ISO is already doing in the context of m-health to foster the quality standards of "apps".

To engage with SDOs to analyse the standards already developed in the context of electronic health records, integrating the work performed by the network on the common data model.

To engage with HL7 and EU Regulatory network to evaluate the possibility of moving from HL7 V3 to HL7 FHIR messaging for ICSRs.

Information technology

Information technology provides the tools for collecting, storing, exchanging, integrating, managing and analysing data from different sources. A growing volume of data produced in different formats by a plethora of heterogeneous sources requires new technologies and architectures to analyse and generate value.

Each new technology has its benefits and drawbacks and the selection of the right tool depends on the individual use cases' requirements. Modern systems often combine multiple data storage and

processing technologies which are chosen based upon the way data is used by individual applications or its components.

Technologies optimised for processing of different data formats supplement traditional relational database management systems. NoSql databases, Hadoop, cloud computing and data lakes are examples of new technologies that provide solutions to storage and processing of large amount of data, which represent a regulatory challenge when exchanging and analysing big data.

These technical solutions must also be supported by proper management processes and data governance.

To foster the use of big data from an information technology point of view it is recommended to:

Implement *bi-modal architecture* to provide space for data exploration and experimentation and turn successful experiments to production.

Embrace *cloud technology* for building big data and analytics infrastructure providing for adequate data transfer speed (or minimising exchange of data).

Implement *security by design*: consideration for *data security and privacy* must precede any adoption of cloud technology.

Develop and implement *data governance processes* supported by the *metadata management.*

Data manipulation

Data manipulation can be defined as the process of *transformation of raw data* to allow patterns in the data to emerge with the aim of *making it ready for analysis*.

Data manipulation is a fundamental step in data analytics; it is often said that 80% of a data analyst's time is spent on this activity. Moreover, the way data is aggregated or recoded can have an influence on the results, either stemming from unintentional mistakes or from being intentional, such as data fishing or the approach used to handle missing data. With increasing sizes of databases, increasing abilities to link data sets and with novel opportunities to collect unstructured data, such as social media data, data manipulation at scale will become an even more important skillset in regulatory authorities.

Considering the time investment made in data manipulation the following is recommended to increase efficiency as well as capability:

Sharing data manipulation approaches that are sound from a regulatory point of view within the network.

Commit to *reproducibility and transparency* by creating a clear framework for reporting and recording the approaches used (e.g. protocolling data manipulation, ensuring adequate capture of meta-data related to an analysis and performing literate coding).

Implement *open source software* to allow the widest possible application of methods across the network and engage with open source software developers to ensure that regulatory needs are incorporated. This will also provide the additional benefit of avoiding the need for some Member States to invest in expensive software.

Artificial intelligence

Analytics is an important tool for gaining insights and providing tailored responses. It can be defined as the research, discovery, and interpretation of patterns within data, and its scientific contribution can be

organised into three classes: i) descriptive, ii) predictive and iii) causal inference. The report focuses on the predictive task which is the field where artificial intelligence (AI) models are applied.

The role of AI is a topic of increasing interest due to the convergence and interrelationship between the vast volume of collectable healthcare *data*, advances in *computational power and storage* and, even if to a lesser degree, advances in *algorithms and methods*.

A framework to provide a more systematic evaluation of developments in this field is described with recommendations: this highlight four areas of *requirements* that AI algorithms should satisfy *for regulatory acceptability*. The four areas encompass being able to produce evidence that is *meaningful* (provides relevance and context sufficient to inform and support decision-making), *valid* (meets scientific and methodological standards), *expedited* (synchronized with the decision-making process) and *transparent* (audible, reproducible, robust, and ultimately trusted by decision makers).

These requirements have a different weight according to the nature of the regulatory decision they will support. Moreover, they have the advantage of highlighting *where attention should be focussed* by either marketing authorisation holders or assessors whilst using or assessing analyses provided through AI algorithms.

Prioritising activities where AI could increase efficiency is recommended, as this objective could *initially* be more acceptable than enhancing support for decision-making. For the latter, validity requirements will play a more important role and concerns about trust and interferences with expert assessment might arise.

To facilitate evaluation and possible adoption of AI initiatives, the following areas should be considered:

Assessing the usefulness of new data and new analytical techniques: it is not necessarily true that larger quantity of data will give more accurate or reliable answers, in particular when the process of data collection is less clear and quality assurance is more complicated. Thus, evaluation against known standards is a prerequisite to adoption of big data and advance analytics.

Piloting: starting small, with methods which are sufficiently transparent and may bring immediate value to the system and creating space for experimentation.

Fostering internal capability: building expertise in data science skills and creating an environment where *subject matter experts work together with data scientists.*

Clustering of expertise: once expertise has been gained through piloting and small initiatives, a dedicated group within a working party and/or different agencies in the network will be beneficial to provide advice, scale the initiatives, and perform horizon scanning. This expert group should also collaborate and gather input from skilled academic groups and external experts.

Characterise advanced algorithms accurately: one barrier to adoption and innovation of AI algorithms is communication. The additional challenge of explaining in human terms results from large and complex models, why a certain decision was reached, is particularly relevant as regulators often want rules and choice criteria to be clearly explainable.

Conclusions

The four topics presented above, and their recommendations highlight *foundational activities* that the regulatory network should engage with, to benefit from big data and the new scientific and technological possibilities. However, the extent of the recommendations proposed require *prioritisation*, *focus*, and a definition on the *mechanism* by which they will be achieved: these activities are the key

deliverables of the second phase of the HMA-EMA big data task force. Nonetheless, one recommendation should clearly be prioritised; independently of whether the regulatory network is main actor in delivering the activities recommended or engages and collaborates with external stakeholders, highly specific skills and knowledge will be required.

Acknowledgments

The authors would like to thank the experts that reviewed and commented the report: Niklas Blomberg and Serena Scollen (European life-sciences Infrastructure for biological Information), Niklas Norén (Upsala Monitoring Centre), Luigi Troiano (University of Sannio), Wo Chang (US National Institute of Standard and Technology), Anja Van Haren (European Network Data Board), Mark Goldammer (Paul Ehrlich Institute), Aldana Rosso (DKMA) and Alison Cave, Jim Slattery and Ralf Herold (EMA).

Data Analytics – Standardisation

1. Standardisation

1.1. Why

The main reason to collect data is always to be able to use it to take informed decisions or to perform reliable, reproducible, scientific and evidence-based assessments.

However, the data journey is usually very complex especially in the Regulatory Medicines Domain; its complexity is also intrinsically linked to the different actions that are performed: collecting, parsing, sharing, integrating, quality assuring, visualising, analysing data and finally interpreting the results of the data analyses.

Standardisation is a fundamental route to overcome these difficulties for the simple reason that standardising the way in which actions are carried is usually the way to "perform" them in a more efficient and effective manner. Standardised processes are the result of brainstorming sessions, discussions, exchange of solutions and experiences between pools of experts. Also, if an action has to be performed for the first time, it is of great help to have as starting point a standard process to follow. Naturally such processes may evolve and improve over time but also may require adaptations to the specific use case before being applied.

For instance, the process of reporting adverse reaction reports (ADRs) in the context of pharmacovigilance provides a useful case study. Currently, in Europe, the process of ADR reporting is described by specific European Regulations (726/2004 as amended and 520/2012 as amended) which reference ISO/HL7 27953 - "Individual case safety reports (ICSRs) in pharmacovigilance". This standard defines the conceptual data model for the information that can be exchanged and the structure of the XML message that has to be used. Standardising the reporting processes results in several advantages; for example, enabling the use of common analytical software and programming scripts which can save significant resources, ensuring Marketing Authorisation Holders are compliant with the legislation and possibly improving the quality of the information reported. Also, having the same data format allows to accelerate data analysis.

In the context of Big Data, following standardised processes is even more important; for instance in processes involving the use of Electronic Health Records for Epidemiological analysis, large volumes of data is required which may come from many independent data sources (e.g. patient health records collected in different country worldwide) and is likely to have different formats which makes integrating the data and analysis extremely challenging.

In general, the benefits of using standards are extensive. Standards help organisations to:

- reduce costs, anticipate technical requirements, increase productive and innovative efficiency,
- augment data sharing, enable the integration of siloed databases,
- increase efficiency and effectiveness,
- enhance data quality and processes,
- foster good analytics,
- ensure that data is collected/provided in a usable and coherent format.

The following case studies can provide further examples:

1. EudraVigilance. This is the European system for the collection of Adverse Drug Reactions; it collects data in a specific data standard and messaging format: ICH E2B (R2). However, with usage, a few weaknesses in the standard were identified and a new **standard data model** and **standard messaging format** (ISO ISO/HL7 27953) was therefore published in 2011. An ICH Implementation guide (referred to as E2B(R3) and a complementary EU ICSR implementation guide were needed to constrain the use of the ISO standard as the ISO standard itself allows some flexibility. The old E2B(R2) message is still in use. EMA and few other stakeholders have implemented the new E2B(R3) message in November 2017 and we are now in a transition period from the old E2B(R2) message. The constrained message format as described by ICH is currently used by WHO-UMC and FDA, Japan and China. Implementation activities are ongoing in other regions like Canada, Switzerland and Brazil. Clearly moving from data collected in silos in various regions using different formats to data still collected in silos but using the same data formats and standards should not only reduce costs for international stakeholders fulfilling similar processes but also foster data exchange and facilitate aggregation of reports across silos to perform signal detection using a bigger set of data.
2. Use of patient health record databases. For many research questions, performing a data analysis on a single database to extract insights and discover correlations will be less powerful than doing the same analysis on multiple databases. For this reason, there is an increasing interest to investigate the same research question and protocol on different patient health care databases. There are 2 possible approaches; the first one is to re-write the programming script such that it can be run on different databases; the second approach is to convert upfront the structures and terminologies of the various databases into a "common data model" to allow exactly the same programming script to run on all databases. The second one is a promising approach especially in the context of a very heterogeneous European data landscape. Also, the second approach help to foster transparency. However, to avoid mapping of data, standardising the data structure of the databases at the point of creation would be preferable but in order to do this, the international community would need to agree and adopt an international standard for the **data model**.
3. Again, utilising patient health records as example but taking an issue even more specific to regulatory decision making, the identification of the drug administered to the patient especially across multiple databases is one of the most time-consuming activity when utilising EHRS. If all data sources adopted the same medicinal product dictionary to standardise the definition of medicine and **terminology**, the return of investment in terms of efficiency, efficacy and data quality would be significant.
4. If two systems or business processes use the same "business concept" and they represent it using different **standard terminologies**, this clearly can create difficulties for those stakeholders that need to utilise both processes. Take for example the coding of indications for medicinal products which can be coded using MedDRA or SNOMED. The need is to calculate the proportion of ADRs received for a specific medicine used for a specific indication (coded in MedDRA in EV) and the number of prescriptions issued for that medicine for the same indication but coded in SNOMED in a patient health record database. The only way to do this is to **map** these two **terminologies** which takes considerable time and resources even for a one-off study and especially if this is to be maintained over time. If instead both EV and the Patient

Health Records system were both coded in MedDRA (for instance), the mapping task would not be needed: this is a clear example of the advantages of **standardising the terminologies**.

1.2. Objectives

Our objectives are:

- Define the concept of standards and standardisation process.
- Clarify ways in which way standardisation can support informed decisions making and reliable, reproducible, scientific and evidence-based assessments.
- Identify standards (already available, under development) that can be helpful in the context of Regulatory Medicines.

Discuss unmet needs for standardisation identified by the Big Data Taskforce Subgroups and propose possible solutions.

1.3. Defining the main concepts

In the context of this report, the overarching term data standard can be defined as 'a model to represent a data entity or series of entities and provides a mechanism to provide consistent meaning to data shared among different information systems.

Every Standard developing organisation (SDO) has its own definition of standards. For ISO, for instance:

"An International Standard provides rules, guidelines or characteristics for activities or for their results, aimed at achieving the optimum degree of order in a given context. It can take many forms. Apart from product standards, other examples include test methods, codes of practice, guideline standards and management systems standards".

In the context of data, we may have standards used for defining data elements, terminologies, unit of measurements, data models, ontologies, data acquisition/collection processes, file formats, electronic messaging and processes related to data analytics.

It has also to be noted that any of the standard types above may be considered a "standard" either because they have been defined via a **formal standardisation process** (e.g. via ISO) or because they are **de-facto globally shared and adopted** or they are **open standards**.

There are different main concepts that can be standardised:

Data element is a unit of data that has a precise meaning or semantic. As such the description of a data element should include a definition, a unit and, where relevant, the process by which the data element was generated.

A **terminology** is a set of "terms" that are shared, unambiguously understood and used among users to represent specific data elements in a database. Examples include SNOMED (Systematic Nomenclature of Medicine)[3], IDC-9 and ICD-10[4], MedDRA (Medical Dictionary for Regulatory Activities)[5]. For instance, in the EudraVigilance system in order to specify the content of the "reaction"

[3] https://www.snomed.org
[4] International Statistical Classification of Diseases and Related Health Problems, 9th and 10th Revision
[5] https://www.meddra.org

data field the MedDRA "terminology" is used. A terminology can be hierarchical or not; MedDRA for instance is hierarchical and has different levels that can be used to describe a medical concept either in a very specific and precise manner (LLT: MedDRA low level term) or in a very broad manner (SOC: System Organ Class). The "Routes of Administration" list used in EudraVigilance is instead an example of non-hierarchical standard terminology.

Measurement terminologies provide a standardisation of units to express "quantities" in the same manner and when not possible (due to different jurisdictions) to have clear and unambiguous unit conversation rules. Universal principles for the expression of measurements have been defined by ISO 31[6], ISO 1000[7] and ISO 80000[8] series of standards, which implement the International System of Units (SI) defined by the General Conference on Weights and Measures. Results of measurements are essential for the identification of medicinal products (strength of the active substance) and other related aspect of the reality (e.g. gene description, laboratory tests results, etc). In addition, it has to be taken into account the ISO 11240 "Health informatics -- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of units of measurement" which is part of the ISO IDMP suite of standards.

A **data model** is an abstract representation which organises data fields in a relational manner to define the relationships between them and to identify how they relate to the characteristics of the real "objects". When this representation becomes widely applied, shared and accepted by stakeholders, it may become a standard data model e.g. ISO/CEN. A data model is made of fields which can be filled using free text, standard and/or measurement terminologies. For example, a data model to define a medicinal product could be composed of a number of data elements including the name of the product, a substance field and its strength, a batch number and a route of administration. A data model is made of fields which can be filled using free text, standard terminologies (e.g. list of routes of administrations) and/or measurement terminologies (e.g. milligrams). For instance, the ISO IDMP suit of standards is now considered the "standard" data model for the unique identification of medicinal products.

An **ontology** is a model that represents within a specific domain artefact with their properties and relationship between them. An ontology purpose is to eliminate conceptual confusion between users and one of the benefits is to speed up information exchange and integration.

A **data acquisition/collection process** is a process in which all the steps for the acquisition and collection of the data (including measurement, storage and validation of the data) are well defined, validated and widely adopted and approved by stakeholders.

An **electronic message** defines an electronic format to exchange a set of data fields in an un-ambiguous and interoperable way between stakeholders. In simple words this represents a way to encode data elements (including sequencing and error handling) to enable the transmission of data from one database to another. An example is the XML message used by Marketing authorisation holders to send medicinal product information to the EMA in the so called "*Art57 database*".

A file format is a standard way to encode data for storage in a computer file. File format are usually specific to the kind of information they store. For instance, a file format "xlsx" is specific to store excel spreadsheets, instead a file format "jpg" is used to store images. This are usually independent from the terminologies but may be incorporated within an overall data standard.

[6] https://www.iso.org/obp/ui/#iso:std:iso:31:en
[7] https://www.iso.org/obp/ui/#iso:std:iso:1000:en
[8] https://www.iso.org/obp/ui/#iso:std:iso:80000:en

Metadata provides information about each dataset, like size, the schema of a database, data format, last modified time, etc. Metadata is normally saved in a common database schema accessible by all users.

1.4. Overview

As already mentioned, standardisation is the outcome of multiple discussions between pools of experts which however need to be formalised, documented, published and maintained by a reliable and independent entity; many "standardisation frameworks" have been established worldwide, but for the scope of this report we will concentrate on the ones relevant for Europe.

In EU, the main standardisation framework is legalised by the European Parliament and Council Regulation (EU) No 1025/2012. Standardisation activities are triggered by market needs and in Europe there are 3 standardisation entities that are in charge of following up with these needs to create the required standards. These are: the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardisation (CENELEC) and the European Telecommunications Standards Institute (ETSI).

However, the standards landscape is extremely complex especially in the world of healthcare: there are many Standard Development Organisations (SDOs) in the World and the European Commission aims to align European standards as much as possible with the international standards adopted by the other recognised SDOs. This process is called "primacy of international standardisation", and it means that European standards should be based on and linked as far as possible with international standards in order to avoid competition between standards or worse conflicting standards. Each European standard adopted as an international standard represents a possible competitive advantage for European industry. To facilitate interaction there are two main co-operative arrangements between the European and international standardisation organisations:

the Vienna Agreement between the International Organisation for Standardisation (ISO) and the European Committee for Standardisation (CEN),

the Dresden Agreements between the International Electrotechnical Commission (IEC) and the European Committee for Electrotechnical Standardization (CENELEC).

Some SDOs are also listed in the Joint Initiative council (JIC). While each member has its own standardisation process to define the types of documents it produces and the mechanism to achieve consensus, the JIC works to enable common, timely health informatics standards by addressing and resolving gaps, overlaps, and counterproductive standardisation efforts. The members of the JIC are:

1. ISO technical committee TC 215 Health informatics (ISO/TC 215): its scope is the standardisation in the field of health informatics, to facilitate capture, interchange and use of health-related data, information, and knowledge to support and enable all aspects of the health system.
2. CEN technical committee TC 251 Health informatics (CEN/TC 251): its scope is the standardisation in the field of Health Information and Communications Technology (ICT) to achieve compatibility and interoperability between independent systems and to enable modularity. This includes requirements on health information structure to support clinical and administrative procedures, technical methods to support interoperable systems as well as requirements regarding safety, security and quality.
3. Health Level Seven International Inc (HL7) : Founded in 1987, Health Level Seven International (HL7) is a not-for-profit, ANSI-accredited standards developing organisation

dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.

4. The Clinical Data Interchange Standards Consortium (CDISC): CDISC is a not for profit organisation and its mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare.
5. SNOMED International: SNOMED International is a not-for-profit organisation that owns, administers and develops SNOMED CT. SNOMED CT is a clinical terminology created by a range of healthcare specialists to support clinical decision-making and analytics in software programs.
6. GS1: GS1 Healthcare is a neutral and open community bringing together all related healthcare stakeholders to lead the successful development and implementation of global GS1 standards enhancing patient safety, operational and supply chain efficiencies.
7. Integrating the Healthcare Enterprise International Inc (IHE): IHE is an initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information. IHE promotes the coordinated use of established standards such as DICOM and HL7 to address specific clinical needs in support of optimal patient care. Systems developed in accordance with IHE communicate with one another better, are easier to implement, and enable care providers to use information more effectively.
8. DICOM (DICOM): DICOM (Digital Imaging and Communications in Medicine) is the international standard to transmit, store, retrieve, print, process, and display medical imaging information.

Additional organisations are:

The International Council for Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use (ICH) is unique in bringing together the regulatory authorities and pharmaceutical industry to discuss scientific and technical aspects of drug registration. ICH's mission is to achieve greater harmonisation worldwide to ensure that safe, effective, and high-quality medicines are developed and registered in the most resource-efficient manner. Harmonisation is achieved through the development of ICH Guidelines via a process of scientific consensus with regulatory and industry experts working side-by-side. Key to the success of this process is the commitment of the ICH regulators to implement the final Guidelines.

International Telecommunication Union (ITU): ITU is the United Nations specialised agency for information and communication technologies – ICTs. Founded in 1865 to facilitate international connectivity in communications networks, we allocate global radio spectrum and satellite orbits, develop the technical standards that ensure networks and technologies seamlessly interconnect, and strive to improve access to ICTs to underserved communities worldwide.

The Big Data Taskforce subgroups' reports have identified standardisation needs for their own business domains. This report captures the needs and where available either identifies possible standards that could be used to satisfy some of these needs or suggests potential new work items to be proposed to specific SDOs or for starting new initiatives outside official SDOs.

1.5. Opportunities (or use) in regulatory activities

Most of the reports produced by the taskforce's subgroups have highlighted needs for standardisation with two solutions:

Identify already existing standards and adopt them or,

Where not available identify the best route to develop a new standard to fulfil the specific business case.

1.5.1. Clinical trial domain

The clinical trial subgroup clearly stated that data standardisation activities are critical to ensure usability and applicability of data. In particular, it is recommended to foster data harmonisation by means of standardising both the data format and the data model being these the basic requirements to be able to combine together data shared from different clinical trials.

In addition, it is recommended the use of harmonised terminology like for instance the International Classification of Diseases (ICD), the Medical Dictionary for Regulatory Activities (MedDRA) and the IDMP standards.

It is important to note that the use of wearable devices has started to be included in the clinical trial process introducing the need of standardisation also in this area.

As mentioned in the subgroup report there are already few initiatives/groups aiming to improve the efficacy and the safety of clinical trials. The main ones are:

ICH (International Council on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use).

CDISC (Clinical Data Interchange Standards Consortium).

CDISC in particular defines an important number of standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. In addition, CDISC standards are already adopted by FDA in the USA, the PMDA in Japan and are endorsed by the CFDA in China.

Opportunity: Europe has not yet adopted a single set of standards for clinical trial submissions. Here, the regulatory opportunity seems obvious: adopt the CDISC set of standards. It is possible that due to the specific EU regulation and guidelines on clinical trials, some of the CDISC standards will need to be amended, but the added value of obtaining a global standard for exchanging clinical trial data is much higher than the effort that will be required for the adaptation. Also, it is important to look in parallel to what other groups, like for instance HL7, are doing in this context to make sure that a possible EU choice is future proofed.

In addition, it is fundamental to be able to unequivocally identify the "substance" that is the subject of the clinical trial for which the finalisation and the adoption of the European Substance Management System (EU-SMS) and the European substance registration system (EU-SRS) will be key. SMS is one of the pillars of the EMA SPOR[9]. These two systems are both based on IDMP. These systems contain different levels of details on substances but will be linked and thus using both it is possible to unequivocally identify a specific substance:

via EU-SMS (SMS is part of the EMA SPOR systems), if it is already used in any other regulatory process (not necessary a clinical trial),

[9] https://spor.ema.europa.eu/sporwi/

via EU-SRS if the same substance is present in other regulatory regions like FDA who use their GSRS system which will be completely linked to the EU-SRS. This allows linkage for instance of different clinical trials in different region using the same substance.

1.5.2. Genomics domain

The genomic subgroup highlighted specifically that while the many different genomics data formats have a similar structure differences arise due to the complexities in:

the different kind of data they contain (sequence, annotations, quantitative data or read alignment),

the need to have specific APIs to exchange data, at the moment there are 3 APIs available developed by different organisation and with overlaps,

the need to standardise the process for the data acquisition.

However, if the accepted data formats for raw data, FASTQ or BAM are used and the data is processed in a standard pipeline, the possibility of sharing the data between databases is ensured.

There are many freely available public databases but currently many are limited by the lack of linkage of genomic data to clinical phenotypes and treatments. Conversely while the genomic information captured through clinical trials by pharmaceutical industry is not freely available it is linked to well curated clinical outcomes; this information could be extremely helpful in various regulatory processes (e.g. pharmacovigilance). To facilitate such linkage across databases, countries and diseases there is a need to standardise genomic data, clinical outcome data and phenotypic data.

However, the developments in genomics and precision medicine together with the diversity of collecting, sharing, coding and exchanging genomic information from various sources has resulted in different terminologies and infrastructures that limit semantic interoperability and data analysis.

Opportunities:

1. evaluate the outcome of EuroGentest[10] and propose its transformation in an ICH or ISO standard,
2. evaluate the work done by different initiatives/organisations like for instance the "Genomic Standards Consortium (GSC)"[11] and Global Alliance for Genomics and Health (GA4GH)[12],
3. create a new standard in ISO or in HL7 for creating an overarching APIs for exchanging genomic data to overcome the limitation of the option to choose between 3 APIs,
4. create a new item proposal in ISO for the creation of Data reporting standard that should combine all the various reporting formats developed so far,
5. consider the feasibility and need of expanding the ISO ICSR standard to include genomic data of the patients,
6. evaluate the possibility of moving from HL7 V3 to HL7 FHIR messaging for ICSRs,
7. there is the need for standardising genomic analysis and data processing techniques,

[10] http://www.eurogentest.org/index.php?id=160
[11] http://gensc.org/
[12] http://GA4GH.org

8. due to the complexity of this field and the different kind of data to be collected and analysed, it would be helpful to create an expert group to propose and lead data standardisation activities. A possibility may be also to include the standardisation objective in the mandate of the pharmacogenomics working party,
9. engage with the Joint Initiative Council (JIC) which coordinates cross-SDO participation in the development of international genomics standards.

1.5.3. Bioanalytics Omics domain

This subgroup highlighted the fact that there are many different data formats and data analysis pipelines. There is a need for (i) standardisation of analytical techniques, and (ii) *"harmonisation of metabolic profiling and biomarker identification for clinical phenotyping in order to meet regulatory requirements".*

Opportunities:

It would be worthwhile to explore if an ISO project could be undertaken to merge all the various initiatives and create a set of ISO Standards in collaboration with CDISC and HL7.

1.5.4. Social media domain

Standardisation in the context of social media is extremely challenging, because the data are mostly free text and unstructured and of very different types across various.

Opportunity: develop a common methodology (API) to retrieve/download data from the various platforms.

In the context of m-Health, there are so many different kinds of apps that standardisation is a challenge. However, what can be standardised is security, safety and privacy. Other topics of importance in the context of standardisation are:

the quality with which the apps have been developed it is an important topic for reliability,

provision of information on the app itself to the user: for instance, who is the developer of the app and which is its affiliation (industry or government),

certification of the app in case this is considered as a medical device.

Opportunity: Engage with ISO TC215 with the specific use case to foster standardisation in the context of mHealth. In ISO TC215 there are specific working groups (WG2: Systems and Device Interoperability and WG4: Security, Safety and Privacy) that are already exploring these topics.

1.5.5. Observational data/Real World evidence (RWE) domain

This report focused on three different types of data sources: Electronic Healthcare Records data and Claims data, Registries, Drug consumption data (Sales and Prescription data).

In general, for all these data sources there are 3 main needs identified:

standardisation of the terminologies used to code information,

standardisation of the data model used with the identification of a "minimum/core data set",

standardisation of the data collection methodologies used.

Indeed, standardised data structure and the use of specific terminologies would improve data collection, quality and data linkage and allow the use of common analytics and coding.

In addition, with regards to registries, the standardisation of the "patient consent" process was recommended.

Opportunity: Engage with ISO to analyse the standards that have been already developed in the context of Electronic Health Records[13]. Also, HL7 may have developed standards in this context. It is possible that those standards can be adopted as they are or that some adaptations would be required. The work already performed on the common data model should be taken into account and integrated with these standards. Another opportunity could be the possibility of proposing a new work item with regards data protection and patient consent in registries. In any case there will be challenges due to the fact that the practical implementation within the MS may be different; also, systems (and terminologies used) may still vary between countries, healthcare domains, regions, or even within organisations.

1.5.6. Spontaneous ADR

This particular source of data is probably the most standardised data set considered through the taskforce with regards to data format, terminologies used and data collection process via the ISO ICSR standard, EU regulations/guidelines and ICH guideline.

The main aspect that could be helpful to standardise is the way of linking genomic data, EHR data and other real-world evidence data with ADRs.

To be noted that the current ISO ICSR is based on HL7 V3 infrastructure which will be decommissioned relatively soon. HL7 FHIR is the new technology adopted in various domains (e.g. ISO IDMP/SPOR).

Opportunity:

engage with ISO and HL7 to consider the feasibility and need of expanding the ISO ICSR standard to include genomic data of the patients,

evaluate the possibility of moving from HL7 V3 to HL7 FHIR messaging for ICSRs,

promote the use of ISO IDMP standards to identify medicinal products or substances as standard in EHR data,

engage with HL7 and EU Regulatory network to develop ISO ICSR message using HL7 FHIR and plan for its adoption.

1.6. Challenges in regulatory activities

Despite the clear opportunities listed above, the use of standards in the EU regulatory domain presents several challenges.

First it must be remembered that the EU regulatory Network is made of many National Competent Authorities and each of them has its own internal national legislations. Therefore, any standard needs to align with multiple national laws. This is challenging because it is almost impossible to have within the SDOs expertise in all the national legal frameworks.

[13] ISO 13606 - Electronic health record communication

Another challenge is related to the fact that a standard may remain completely neglected if it is not referenced in, for instance, an EU Regulation. Indeed, sometime adopting a new standard may require important investments with gains that materialise only after few years and also only if a high percentage of the relevant stakeholders implement it. A downside of enforcing an IT standard via an EU Regulation is losing flexibility. The clinical domain IT-sector evolves with a much faster speed than the regulatory domain and there is a risk of being one step behind.

A further important challenge is related to timelines; in order to publish a standard, many months (if not years) may be required and this often may discourage the initiation of new projects. This is particularly pertinent for fast developing innovation, for instance the use of Artificial Intelligence to support regulatory processes, these timelines of publishing a standard may mean that by the time a standard is published the landscape may already have evolved to such an extent that the standard is already obsolete. Or implementation activities within the regulatory landscape take so long that the outside world (clinical domain) has already moved further by the time the regulators are finally ready.

As already mentioned, the standard development process is based on consultation and collaboration between worldwide subject matter experts. In order to develop a standard significant human and financial resources are required and the return of investment must be sufficient to outweigh these costs. It is this important that the anticipated benefits of using the proposed standard are greater than the actual developing costs (ROI analysis[14]). The benefits should include not only potential savings related to human and financial resources, but also aspects such as the possible benefits for public health. An evaluation process for the ROI should be developed and implemented within any projects initiated.

It also critical that all EU Regulators participate in the development of relevant standards in order to ensure that standards fulfil their national needs (both from the regulatory and clinical domain). Often there is a disconnection between Regulators and SDOs and this vacuum should be filled. The European Network Data Board[15] has in its mandated the objective of "proposing common policies, procedures, architecture and **standards** to maximise the sharing and investment in data and information" and therefore it can contribute to fill this vacuum when relevant standardisation topics are brought in the Agenda; latest example is relevant to the implementation of the ISO IDMP standards.

It is also important to make a distinction between:

Standards that are directly applicable to the regulatory domain (e.g. ISO IDMP and ISO ICSR),

Standards hat have a primary use case in the clinical domain and the regulatory domain as secondary use case.

Based on this distinction, the opportunities may be different (e.g. for influencing the content, but also for enforcing a standard through legislation) and the level of engagement of regulators can also vary.

AI is gaining increasing prominence in not only academic science but also in the popular press and recent concerns have been raised in the ISO JTC1/SC42 Artificial intelligence (and big data) group related to the trustworthiness of solutions using AI. They have created a specific working group to deal with this aspect which is one of the most challenging one that the regulatory domain will have to face in the months and years to come. For instance, how will regulators be able to accept as valid, evidences derived from AI algorithms when used as support for a Marketing Authorisation application?

[14] https://en.wikipedia.org/wiki/Return_on_investment
[15] https://www.ema.europa.eu/en/documents/other/european-union-network-data-board-terms-reference_en.pdf

It will be extremely important that regulators participate in these ISO discussions to articulate the Regulatory business needs for protecting patient health.

Standards from ISO, CEN are copyrighted, and this may limit their use where not mandated by legislation: stakeholders may decide to look first at freely available alternatives (open standards) before purchasing ISO/CEN standards. Other SDOs for instance have different **financial models** which allow them to give access to the standards (or part of them) for free to stakeholders (e.g. HL7 FHIR, etc). For Health-related standards EU should have centralised funding from EC, NCAs and research.

1.7. Regulatory implications

All subgroups of the HMA/EMA Big data taskforce highlighted standardisation as the primary element on the road from big data to regulatory acceptability, prominent support for standardisation from within the regulated medicine domain will drive significant benefits. For instance, it can improve:

a) **data collection**: the availability of standardised format and processes will enhance the quality of the data, the speed with which the data is collected and the consistency of data,

b) **data quality**: the availability of standards defining data quality principles will facilitate the creation of data quality frameworks within processes and IT application adopted or developed within the Regulatory Network. For instance, ISO IEC JTC1 SC7 has developed two standards ISO/IEC 25012 "Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model" and ISO/IEC 25024 "Systems and software Quality Requirements and Evaluation (SQuaRE) -- Measurement of data quality" that can be relevant in this context,

c) **data sharing:** the availability of standardised formats and terminologies will facilitate the linkage of data in an easier and more timely manner,

d) **trustworthiness:** If specific AI algorithms are validated and accepted by an international and formal working group (e.g. in ISO), they will be of higher data quality and therefore the risk of errors will be reduced,

e) **regulatory acceptability:** Algorithms built following standardised principles or validated by international and formal fora will be more acceptable for regulatory decision making,

f) **global harmonisation:** The adoption or developing of standards as a worldwide effort will facilitate harmonisation between the regulatory processes worldwide. In the medium-long term period (after national adaptations) regulators will benefit from globally adopted processed, standardised terminology by easier data exchange, the possibility to integrate multiple national data sources and common data interpretation following data analyses,

g) **governance:** The dedicated working group on "AI governance implications for organisations" created within the ISO SC42 may help EU regulators to avoid errors in this important field.

1.8. Conclusions

Standardisation across multiple fields is crucial to gain the insights that 'big data' can offer and to adopt dedicated AI techniques in the regulatory domain. Standardisation may speed up implementations of processes and analytical systems and foster harmonisation that in return will widen data analytic possibilities for regulators.

Therefore, the key messages are:

to engage with Standard Development Organisations (e.g. CEN) and other organisations that are developing open standards (e.g. GA4GH),

to ensure that legislators incorporate where possible agreed standards when developing guidelines and regulations,

to adopt (and adapt when needed) international standards to foster harmonisation and improve efficiency, if the ROI analysis presents a positive benefit,

the European Network Data Board (EUNDB), consulting with other scientific and IT committees and working parties, should remain the forum for the EU regulators to raise and discuss standardisation needs, to develop consensus positions on the implementation of new standards fora (e.g. ISO plenary meetings),

if not already in place, National Regulators should also initiate direct partnerships with their national standardisation bodies and EMA should develop similar partnerships with European wide bodies such as CEN.

1.9. Recommendations

The overall core recommendation is to "promote the use of global, harmonised and comprehensive standards to facilitate interoperability of data" and the following principles should apply:

Minimise the number of standards; strongly support the use of available global data standards or the development of new standards in fields where none are available to ensure early alignment,

Where data cannot be standardised at inception, establish the regulatory requirements to confirm the validity of mapped data,

Promote use of global open source file formats.

Further detailed recommendations on data standardisation are provided below.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
1	Governance	Bridge the gap between National standardisation bodies and National competent authorities	Promote the EU Regulatory Network standardisation needs (including the ones identified by the HMA/EMA Big data taskforce) at the European Network Data Board and achieve consensus on how to address them. When a need for a new project relevant to regulatory use cases is identified by any standard development organisation (SDO), the European Network Data Board (EUNDB) representatives should seek the opinion of their national SDOs to ensure alignment.	ISO national organisations aware and aligned with the views of the National Regulators
2	Governance	European Network Data Board (EUNDB) should actively engage in activities related to standardisation that can impact the work of the Regulatory Network	With regards to 'omics data: involve the pharmacogenomics working party to propose and lead data standardisation activities in collaboration with the EUNDB. Continue and strengthen the communications between international Regulator authorities to ensure harmonisation of standardisation approaches.	Increased engagement of EUNDB with EMA scientific working parties
3	Governance	Engage with EC	Once a new EU legislative proposal regarding collection and use of specific data in the regulatory domain is under development, a stronger engagement with EC should be implemented to ensure that standards proposed by the experts of the EU Regulatory Network are included in the legislation. Contribute to the European approach to artificial intelligence and robotics by means of engaging with the EC High-Level Expert Group on Artificial Intelligence to bring forward the use cases relevant to Regulated Medicines.	Legislators supporting the standardisation needs of the EU Regulatory Network Open dialog with the EC High-Level Expert Group on AI in place

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
			For Health-related standards EU should have centralised funding from EC, NCAs and research.	
4	Data Management	Data Mapping	Where data cannot be standardised at inception, establish the regulatory requirements to confirm the validity of mapped data for the proposed application. Continue and if necessary, strengthen engagement with the IMI EHDEN project. If a data mapping is proposed, the process and financial model to maintain the mapping over time should be defined at the outset of the project.	Developed Data mapping frameworks. Increased engagement with ad-hoc projects focusing on data mapping.
5	Metadata Management	Maintenance of metadata information to facilitate data linkage	Create over-time a metadata repository for all the databases used by the EU Regulatory Network. When creating a common data model, it is critical to include also a common metadata repository.	Increased use of metadata repositories
6	Governance	Standardisation process	Create a framework to ensure that the anticipated benefits of using or developing a standard are greater than the actual developing costs. The framework should incorporate include not only potential savings related to human and financial resources, but also aspects such as the possible benefits for public health. Before a project is started a Return of Investment (ROI) analysis is recommended.	Steps for benefit/costs evaluation included in the project's pre-inception phase
7	Engagement with Stakeholders	Engagement with organisations that collects or generates the source data	The EU Regulatory Network should create a framework and communication channels to engage with data provider and	Decrease in the number of terminologies used by different data providers

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
			data generator organisation (e.g. industry) to ensure that they align to the data standards desired by regulators	
8	Engagement with SDOs	AI and Analytics standardisation	Continue to engage with ISO SC42- "Artificial intelligence and in particular with its working and studying groups: Governance implications of AI Computational approaches and characteristics of artificial intelligence systems AI foundational standards Trustworthiness Use cases and applications Big Data. Participate to the ISO SC42 "Artificial intelligence" work items on: "Governance implications of the use of artificial intelligence by organizations" "Bias in AI systems and AI aided decision making" "Big data reference architecture".	EU contribution to the SC42 work item provided
9	Engagement with SDOs	Engage with CDISC to foster CT data	Strongly support the adoption in EU of CDISC data formats, terminologies and messaging in combination with other relevant terminologies (e.g. ICD, MedDRA, ISO IDMP). Given the endorsement by a number of international regulatory bodies, CDISC is becoming the de facto SDO for clinical trial data. Adopting CDISC standards could ensure alignment of EU Regulatory Network with other regulators and facilitate global clinical trial data sharing. At the same time, explore what is being developed in other groups, like for instance HL7, to make sure that the choice of CDISC is future proofed.	Improved clinical trial data harmonisation

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
10	Engagement with SDOs	Engage with ICH, ISO, HL7 and other genomic data standard initiatives such as the Global Alliance for Genomics and Health (GA4GH)[16] to drive standardisation of 'omics data	Propose a new project to convert and further develop EuroGentest to a possible standard and if relevant for incorporating genetic testing for medicinal diagnosis. Propose a new project to create an overarching APIs for exchanging genomic data to improve harmonisation. Propose a new project proposal aiming to create a data reporting standard that should combines all the various reporting formats developed to date. Propose a new project to standardise genomic analysis and data processing techniques.	Improved harmonisation and data sharing capabilities related to 'omics data
11	Engagement with SDOs	Engage with ICH, ISO or HL7 to foster data standards to be used in the context of Bio-analytics Omics	An ISO project to merge all the various initiatives and create a set of ISO Standards in collaboration with CDISC and HL7 would bring significant value to this field by means of creating common data format and terminologies. Such standards should equally apply to metadata. Collaborate with ISO SC42 to standardised linkage protocols and techniques based on AI for aggregating together different datasets.	Improved harmonisation of the data formats and terminologies
12	Engagement with SDOs	Engage with ICH or ISO in the context of social media and m-health	Engage with ISO TC215 to review the current standards related to mHealth apps and determine if new work item proposals are required.	Improved data sharing capabilities related to social media and m-Health data

[16] http://oicr.on.ca/oicr-programs-and- platforms/global-alliance-genomics-and-health-ga4gh

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
			Propose a new work item to develop a common methodology (API) to retrieve/download data from the various social media platforms.	
13	Engagement with SDOs	Engage with ICH or ISO and EC in the context of observational data	Engage with ISO and EC to analyse the standards that have been already developed in the context of Electronic Health Records. This in particular will be for standardisation of the terminologies used to code information: standardise the data model used with the identification of a "minimum/core data set" following the proposal of the EMA registry initiative standardisation of the data collection methodologies used Adapt and adopt (via new work item proposals) the identified standards.	Increased harmonisation related to the data format and content of observational data
14	Engagement with SDOs	Engage with ICH or ISO in the context of ADR data	Engage with ISO and HL7 to consider the feasibility and need of expanding the ISO ICSR standard to include genomic data of the patients. Evaluate the possibility of moving from HL7 V3 to HL7 FHIR messaging for ICSRs. Drive the adoption of ISO IDMP standards in EHR data, disease registries, drug utilisation registries, cohort studies and other relevant data sources to identify medicinal products or substances. Maintain a full mapping of the terminologies used for indication of medicinal product and adverse drug reactions between ADRs and HER.	Improved quality and precision of ADR reporting Improved linkage capabilities between ADR and HER data, disease registries, drug utilisation registries, cohort studies and other relevant data sources

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
15	Engagement with SDOs	Foster standardisation in EU	When open data standards are not possible reach agreement with the specific SDO to allow Regulators to use parts of the copyrighted standards in guidelines.	Improved integration of standards with Regulatory guidelines
16	Engagement with SDOs	Engage with International Telecommunication Union (ITU) on AI for health	Closely monitor the work delivered via the ITU/WHO forum and engage where working items are relevant for the EU regulatory network.	Input provided in the work items produced by ITU and relevant for the EU regulatory network
17	Engagement with SDOs	Engage with ISO	Establish a mechanism to exchange huge amount of data between organisations. At the moment exchanging 1 TB of data via network may be a challenge. Define standard specialised compression algorithms to reduce the size of the data and define new network protocol more suitable to exchange this huge amount of data.	Data Exchange facilitated
18	AI specific topic	Trustworthiness	Regulators to attend the ISO SC42 meeting discussions with regards to trustworthiness and regulatory acceptability to bring forward the EU Regulatory Network business needs for protecting patient health.	Input provided to the ISO discussions on trustworthiness
19	Mapping terminologies	Expand/utilise the HL7 BRIDG project to EU data sources	Active participation in the HL7 BRIDG project within the BR&R working group. Identify the data elements within the BRIDG model that are relevant for EU. Identify the data elements relevant for EU but missing in the BRIDG model and try to harmonise the regions to reach global interoperability for EHR data.	Bridge model extended to be fit for purpose for EU use

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
20	Terminologies	Adopt ISO IDMP for the unique identification of medicinal products and substances to foster data linkage	Utilise the Product Management Systems (PMS) for the unique identification of medicinal products in all possible use cases. Utilise the Substance Management Systems (SMS) and the EU Substance Registration System (EU-SRS) for the unique identification of substances in all possible use cases.	Improved data linkage between different data source using the substance or product identifiers as defined in PMS, SMS or EU-SRS

1.9.1. Subgroups recommendations supporting the needs or standardisation

The various subgroups have expressed several needs in their reports, and these have originated some recommendations specific for their areas. Those are reported below for reference including the original reference number as in the "HMA-EMA Joint Big Data Taskforce – Summary report".

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
1	Data standardisation activities are critical to increase data interoperability and facilitate data sharing	Agree on data formats and standards for regulatory submissions of raw patient data	Strongly support the use of available global data standards and alignment with other regulatory bodies to facilitate global clinical data sharing e.g. CDISC and ISO IDMP. Encourage use of open source data formats global standards. Establish guidelines for use of other types of data types such as DICOM for images (see recommendation no. 5) relevant to regulatory submissions.	Agreement on formats and standards
6	Inconsistent availability of healthcare data from secondary care	Mechanisms are required to drive the, standardisation and access to secondary care data	Proactively support approaches to improve the linkage of primary and secondary healthcare data. Proactively support pilots for areas where data is lacking e.g. linkage of paediatric data across specialist paediatric centres. Encourage standardisation of terminologies across care settings to facilitate linkage. • Create an inventory of reliable sources of secondary care data hosted on the ENCePP website.	Increased access to data from secondary care
8	Timely access to pan European healthcare data	Sustainable mechanisms for combining healthcare data across Europe should be implemented	Strongly support the establishment of distributed networks of datasets to improve timely access to data. Where networks utilise a Common Data Model:	The speed of real-world evidence generation across multiple datasets

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
			– ensure the impact of transformation of data on the evidence generated is understood; – define the regulatory use cases for which distributed data networks dependent on a CDM would be acceptable; – ensure the maintenance of up-to date mappings as new data elements are introduced. Support the development of robust data governance mechanisms to ensure data privacy obligations. Emphasise the need for sustainable solutions.	
9	Multiple coding systems to record exposure and outcomes from medicinal products	Increase the consistency of recording information on exposure to medicines including indications for use, product, dose and route, duration Increase the consistency of recording of outcomes	Support the implementation of ISO IDMP standards within electronic health records. Support the mandatory recording of indications. Support the mandatory recording of outcome measures including cause of death.	Increase in the consistency in the recording of exposure to medicines and utility of RWD
11	Improve the integration of new data sources	Mechanisms should be developed to integrate new data sources with EHRs	Support the development of standard terminologies and methodologies to enable the incorporation of data from novel data sources e.g. m-health, PROM in a consistent manner.	Increase in the availability of consistent information of lifestyle factors and PROMs in EHRs

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
12	Implementation of common core data elements in registries Specify data quality attributes for data standards	Harmonisation of data elements, standards, terminologies and quality attributes to improve data interoperability	Facilitate agreement by registry holders on common core data elements to be collected by all registries in a given disease area. Contribute and support the definition and inclusion of data elements relevance for medicines evaluation e.g. ADRs, co-morbidities. Where possible common data standards and coding systems should be used. Establish minimum set of data quality attributes acceptable for regulatory purposes across multiple disease areas.	Publicly accessible list of the common data elements (with their definitions) collected by registries in a given disease area Increased use of registries in regulatory submissions
16	Multiple coding systems to record exposure to medicinal products	Increase the consistency of recording information on exposure to medicines including product, dose and route	Support the implementation of ISO IDMP standards. Implement mandatory recording of indications for use.	Increased consistency in the recording of exposure to medicines Reliable verifiable linkage with community dispensing records
32	Pharmaco-epidemiology studies	Promote the use of m-Health technology to support effective post-authorisation studies	Support case studies of m-Health technologies in order to better understand how these technologies could increase the strength of post-authorisation studies. Establish standards for consistent data collection across apps.	An increase in the use and value of m-Health data within post-authorisation research

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
34	Standardisation and data linkage	Optimise data sharing and linkage of phenotypic and/or treatment parameters to genomics datasets	Promote the use of harmonised open data file formats to improve sharing of genomics data and/or clinical outcome data linked to genomics data. Promote linkage of relevant parameters (e.g. adverse events, primary efficacy outcomes) to the genomics dataset upon marketing authorisation application. Promote interoperability of genomics data platforms.	Increased linkage of genomic data to the key clinical parameters
44	Supporting the harmonisation of data (file) formats	Harmonisation of the used data (file) formats	In order to establish an Open Data Mandate, it is crucial to identify or develop open source file formats which include the relevant data and information (e.g. relevant metadata). Regulatory agencies should advise which data file formats and / or attributes of data formats are acceptable for regulatory purpose.	Increase in the number of available, relevant and harmonised 'omics' big data sets acceptable for regulatory decision making
45	Strengthening the development and harmonisation of data standards	It is encouraged to minimise the number of data standards used	Suitable and appropriate data standards should be identified and if necessary, adapted for the use in big data approaches. Data standards should be platform-independent, appropriately validated and freely available.	Increase in the number of available, relevant and harmonised 'omics' Big Data sets acceptable for regulatory decision making

1.9.2. Useful references

ISO SC42

Standardization in the area of Artificial Intelligence. Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence. Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications.

Big data Value Association

The Big Data Value Association (BDVA) is an industry-driven international not–for-profit organisation with 200 members all over Europe and a well-balanced composition of large, small, and medium-sized industries as well as research and user organizations. BDVA is the private counterpart to the EU Commission to implement the Big Data Value PPP program. BDVA and the Big Data Value PPP pursue a common shared vision of positioning Europe as the world leader in the creation of Big Data Value.

AI Alliance

Given the scale of the challenge associated with AI, the full mobilisation of a diverse set of participants, including businesses, consumer organisations, trade unions, and other representatives of civil society bodies is essential. The European AI Alliance will form a broad multi-stakeholder platform which will complement and support the work of the AI High Level Expert Group in particular in preparing draft AI ethics guidelines and ensuring competitiveness of the European Region in the burgeoning field of Artificial Intelligence.

High-level Expert group on Artificial Intelligence

The European Commission has appointed 52 experts to a new High-Level Expert Group on Artificial Intelligence, comprising representatives from academia, civil society, as well as industry. The High-Level Expert Group on Artificial Intelligence (AI HLEG) will have as a general objective to support the implementation of the European strategy on Artificial Intelligence. This will include the elaboration of recommendations on future-related policy development and on ethical, legal and societal issues related to AI, including socio-economic challenges. Moreover, the AI HLEG will serve as the steering group for the European AI Alliance's work, interact with other initiatives, help stimulate a multi-stakeholder dialogue, gather participants' views and reflect them in its analysis and reports.

ITU AI4Health

Aimed at leveraging the power of artificial intelligence (AI) to advance well-being and quality health care for all, the ITU Focus Group on AI for Health brings together specialists to develop a benchmarking framework for international standards and steer the creation of policies to ensure the safe, appropriate use of AI in the health sector and identify use cases that can be brought to a global scale. AI-powered health data analytics can enhance medical diagnostics and improve decision-making about treatment options and health interventions. With over 1.3 billion people owning a smart phone, AI based solutions can bring diagnostics to people with limited or no access to medical care, free up capacities of health professionals to focus on critical cases, or help save lives in emergencies through allowing patients to be diagnosed even before they arrive in hospitals to be treated.

IEEE Big Data

Big data is much more than just data bits and bytes on one side and processing on the other. It entails collecting, storing, processing, and analysing immense quantities of data that is diverse in structure in order to produce insights that are actionable and value-added. Vast amounts of data of various types are being generated at increasing rates. Determining how to utilize this data strategically and efficiently is the goal of technologies associated with the Big Data initiative. Merely collecting and storing data is not the sole objective of Big Data; rather, enhancement of businesses or societies drives the technologies of Big Data. For example, successful big data solutions can provide targeted marketing, identify new markets, or improve customer service through analysis of customer data, social media, or search engine data. Examination of industrial sensor data or business process data can enhance production, aid in proactive improvements to processes, or optimize supply chain systems. As a final example, society can benefit from big data analytics through intelligent healthcare monitoring, cybersecurity efforts, and smart cities data manipulation.

IEEE Big Data Governance and Metadata Management (BDGMM)

Governance and metadata management poses unique challenges with regard to the Big Data paradigm shift. The governance lifecycle needs to be sustainable from creation, maintenance, depreciation, archiving, and deletion due to volume, velocity, and variety of big data changes and can be accumulated whether the data is at rest, in motion, or in transactions. Furthermore, metadata management must also consider the issues of security and privacy at the individual, organizational, and national levels. From the new global Internet, Big Data economy opportunity in Internet of Things, Smart Cities, and other emerging technical and market trends, it is critical to have a standard reference architecture for Big Data Governance and Metadata Management that is scalable and can enable the Findability, Accessibility, Interoperability, and Reusability between heterogeneous datasets from various domains without worrying about data source and structure. The goal of this Activity is to enable data integration/mashup among heterogeneous datasets from diversified domain repositories and make data discoverable, accessible, and usable through a machine readable and actionable standard data infrastructure.

NIST Big Data Public Working Group

Develop a secured reference architecture that is vendor-neutral, technology- and infrastructure-agnostic to enable any stakeholders (data scientists, researchers, etc.) to perform analytics processing for their given data sources without worrying about the underlying computing environment.

Data Analytics - Information Technology for Big Data

2. Information Technology

2.1. Why

Modern world produces enormous amounts of data coming from various sources and often available through internet. This richness of data often called "big data" can be analysed computationally to reveal patterns, trends, and associations that can help organizations make informed decisions.

Information technology is an enabler for big data analytics. It provides the tools for acquiring, storing, moving, transforming, integrating, processing, analysing and managing large amounts of data coming from different sources. A growing volume of complex data produced at high rate in different formats by a plethora of heterogeneous sources requires new technologies and architectures to analyse and generate value from all this data.

Each new technology has its benefits and drawbacks and the selection of the right solution very much depends on the individual use cases' requirements. Modern systems often combine multiple data storage and processing strategies, often designed on specific needs of individual applications but still applicable to other contexts and scenarios.

2.2. Objectives

This chapter focuses on the landscape of modern technologies for big data management and analytics, attempting to provide their general overview. Thus, the chapter does not go deeply into technical details but tries to describe their most important aspects aiming to guide selection of right tools and addressing to technical literature for further information.

2.3. Main concepts

2.3.1. Big data

Big data is a term used to refer to data sets that often due to their volume and complexity are not easily managed by conventional databases and data processing systems. Big data is often characterised by 5Vs[17]:

Volume refers to vast amounts of data generated every second, such like data generated from sensors or data generated by internet users. These amounts are often too big to store and analyse using traditional database technology.

Variety refers to different types of data captured – structured, semi-structured, unstructured, pictures, video, voice. Traditional data base technology was created to work with well- structured and normalised data and has problems processing other data structures.

[17] Volume, Variety and Velocity have been originally introduced by Doug Laney in a report "3D data management: Controlling data volume, variety and velocity" published by Gartner in 2001. Later, several other "V" properties have been presented in literature in order to characterize big data, of which most accepted are Value and Veracity.

Velocity refers to speed at which data is generated and at which it needs to be moved around. For example, sensor generated data needs to be captured in real time otherwise it gets lost. Traditional data processes are not performant enough to cope with this speed.

Veracity refers to accuracy and trustworthiness of data. Some big data sources are not providing high quality data or may even contain purposefully falsified information (e.g. fake news in tweets). Often volumes of data make up for the lack of quality. The latter concept is strongly disputed in case of analysing clinical data, where a small, underpowered but controlled dataset is often preferred to a large amount of uncontrolled data.

Value refers to ability of turning big data into business value for an organisation that balances investments in big data infrastructure.

What makes big data different from conventional data is the fact that attempting to provide an analytical answer we cannot escape from considering the complexity, amount and speed of data as part of the solution.

2.3.2. Big data sources

Big data is everywhere, and the data has become so complex and so dynamic that it can provide great insights, but what are the various sources of big data?

Enterprise data owned by organisations is still the primary source of big data. It includes data stored in databases, document repositories, emails and even file systems. Much of enterprise data is still untapped for data analytics.

Publicly available data sources and open data available over internet are another important source of data.

Sensor data produced by both specialised medicinal devices and commercially available fitness trackers can provide another source of valuable data for the healthcare.

Web generated **data about transactions** and behaviour of users is another important source. This can include information about web searches and visited web pages of internet users, history of their purchases, etc.

Ontologies and taxonomies organising data can be a source of big data as they already include carefully organised relationships between data. Such medical taxonomies include MedDRA, ATC or SNOMED.

Social media are a source of very raw, dynamic and diverse data in many formats, incl. free text, videos, sound and graphic.

Figure 1: Big data sources

2.3.3. Big data formats

Data can be stored in different formats, varying from well-structured records to unstructured free text documents and pictures. Each format requires different methods of storage and processing.

Structured data have a high degree of organization following a predefined data model, with data properties and relations known a priori. Relational models are very common to describe structured data and several formalisms have been proposed over the time to do so (e.g. ER and UML notations). A common way to represent structured data is by means of linked data tables, where each table column represents data property and each table record is a data point. Relations can be created between such tables. The technology of relational data base management systems (RDMBS) is well established and it has been largely developed by several vendors since 70s.

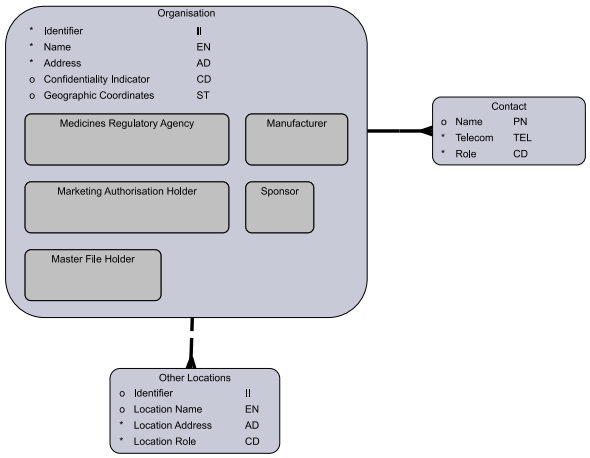


Figure 2: Sample structured relational data model

Spreadsheets (e.g. in Excel) under some circumstances can also be treated as structured data. Theoretically, spreadsheets are relational tables but because there are no restrictions on the method of input, they can easily turn into semi structured data that needs to be normalized for machine reading.

Unstructured data is not organized into a predefine format. It more resembles the way actual humans communicate and think. Examples of unstructured data are text documents, PDF files,

PowerPoint presentations, emails, etc. This kind of data is difficult to interpret and process by automatic systems as its structure may vary, data may include discrepancies, omissions, misspellings and other issues easily resolved by people but difficult for traditional computer programs.

> Given that unstructured information will account for 90% of all data created over the next decade[18], it is no surprise that companies are investing in NLP (Natural Language Processing), artificial intelligence (AI) and data mining technologies in order to tap into, and one day fully interpret, this well of data.

Semi structured data is somewhere in the middle between structured and unstructured data. It very much resembles an electronic version of a paper form, with a sequence of fields mixed with free text fragments. Each form represents an individual document and each document roughly follows the same format. Each semi-structured document is typically treated, stored and processed as a single entity and not as a collection of individual database tables known from traditional databases. The most popular examples of semi-structured data formats are documents expressed in XML or JSON.

Figure 3: Sample semi-structured JSON document

Digital media data formats include video, sound and graphics. These are very special cases of unstructured data. Understanding the content of this kind of data has been traditionally a challenge for computer systems, but recent developments in artificial intelligence for speech recognition and computer vision are changing it. For example, there are already use-cases showing the possibility of computers in assisting or even replacing specialists in medical imaging analysis[19].

2.3.4. Data analytics models

Data analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Please refer to the Data Analytics chapter for more information on this topic.

Traditional approach to data analytics was controlled and centralised, depending much on pre-defined, descriptive reports for business to analyse.

This approach proved to be too static for the users, who wanted more self-service ability to gain new insights into the data. Initially self-service provided the business users freedom of reporting, by

[18] Source: https://www.computerworld.com/article/2509588/world-s-data-will-grow-by-50x-in-next-decade--idc-study-predicts.html, accessed 4 Jun 2019
[19] Source: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ai-diagnostics-move-into-the-clinic, accessed 4 Aug 2018

allowing them to build and customize their own reports without IT intervention. These reports were typically created based on well-understood data and pre-cleansed sources.

Some users still want to access raw data from completely new sources and explore it on their own to discover unforeseen dependencies. This is much more rapid and agile approach to data analytics. Users want to experiment with data and understand that some of these experiments will initially not produce insights.

New trends take self-service further, by offering them freedom of the actual data access, in which business users can obtain any set of data in the organisation without IT intervention. Experimentation with data needs to be done quickly and relatively cheaply, as experimental activities by their nature may not lead to any valuable results.

Gartner proposed a three-tier model for data analytics[20]:

The **analytics portal** is targeted in providing traditional descriptive reports based on trusted and structured sources. Reports are mainly developed by IT.

The **analytics workbench** offers more freedom to business to provide more agile insights into data and generate descriptive and simple diagnostics reports. The reports can be built by business power users without IT involvement.

The **analytics data science laboratory** is aimed at advance analytics resulting in complex diagnostic analyses, predictive and prescriptive reports. Data scientists, business users with deep understanding of data sources and analytical methodologies can use this lab for data exploration and experimentation.

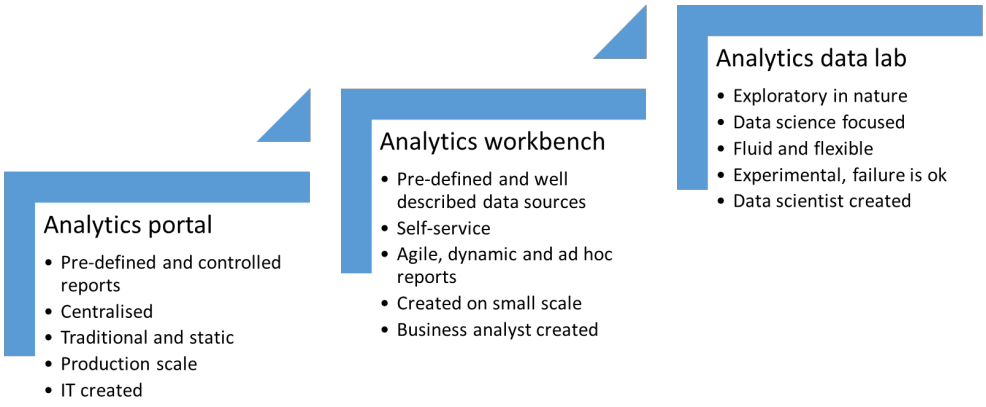


Figure 4: Three-tier model for data analytics

2.4. Overview

2.4.1. Data storage technologies

Different data formats require adequate technology for storage and processing. Technologies optimised for processing of different data formats supplement traditional relational database management

[20] "2017 Planning Guide for Data and Analytics", Gartner, 13 October 2016

systems. A new trend is to combine multiple data storage technologies[21], chosen based upon the way data is used by individual applications or components of a single application.

Relational databases implement well-structured relational data models. These databases use SQL - a standard data definition and query language and are often referred to as **SQL databases**. The relational databases gained huge popularity in 1980s and are still widely used. They are a well-known and understood technology with support from a multitude of SQL based tools.

Relational data is organised into well-defined tables with data columns of a fixed datatype and maximal length. Primary keys are used to identify individual records within each table, and foreign keys are used to link records in different tables as defined by relationships defined in the model.

Relational databases traditionally were optimised to process a large number of real-time transactions, with multiple data records being written and read in parallel (e.g. tracking financial transactions in banking systems). Recently there is a raise of more specialised SQL databases optimised for other scenarios (e.g. quick responses to SQL queries over very large data sets, but relatively little frequency of data updates). The selection of the right SQL database depends now very much on a particular use case and its requirements.

NoSQL databases address the main disadvantage of relational databases – their fixed data model that is relatively difficult to evolve for existing applications. NoSQL term is used in opposition to relational databases that implement the SQL standard. There are many very different products calling themselves NoSQL and there are no universal standards. Each NoSQL database comes with own strong and weak points, and the selection of the right tool very much depends on particular requirements. Some of the differences between these products can be very technical, like properties of transaction processing or ability to keep data consistency in a distributed database in case of network disruptions.

Regardless of their technical differences, it is generally agreed that NoSQL databases can be categorised according to the type of data structures they can process:

Key-value databases are the simplest form of NoSQL data stores. They represent data as pairs: a key linked to a value. In pure key-value stores, the database can only interpret and search the key, whereas the value is opaque, i.e. the database cannot interpret its content. Value can be accessed only based on the related key. It is up to application built on top of key-value database to interpret and process this value. These databases are extremely fast, both for read and write operations, as they only work with a key and just read/ write the value. But this model limits their use only to particular use cases. They are ideally suited for scenarios where persistent caching is required (e.g. on-line shopping cart based on customer's id used as a key and cart content as a value), storing web application session information or user profiles and preferences. Sample implementations include: Redis, Dynamo, Riak, Oracle NoSQL Database.

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

[21] Polyglot persistence is a term often to describe such scenarios when storing data, it is best to use multiple data storage technologies, chosen based upon the way data is being used by individual applications or components of a single application.

Figure 5: Key-value database visualisation. Values can be anything[22]

Document databases organise data by storying a unique key linked to a document, which content can be interpreted by the database. In a sense they are extension of key-value databases, where value (i.e. document) has structure that can be accessed and queried by the database. It is important to note that documents in this context are structured or semi-structured files, typically in JSON or XML format, and not arbitrary unstructured binary documents in PDF or graphical format. Each document can have its own structure, with deep nested hierarchies. This gives much flexibility to store a variety of data formats. It makes document databases well suited for applications that store data in flat or structured files or for data that has structure too complex or diverse for relational databases. For example, this data model would be suitable for storying structured medical forms in JSON format. Sample implementations include: MongoDB, CouchDB, Jackrabbit.

Key	Document
employee_001	{ "firstName": "John", "lastName": "Smith", "age": 27 }
employee_002	{ "firstName": "Mary", "lastName": "Smith" }
employee_003	{ "firstName": "Robert", "lastName": "Newton", "age": 66 }

Figure 6: Sample document database with JSON documents

Column databases organise data in rows and columns. A row can have a variable number of columns, possibly even thousands of columns. Each row is identified by a row key, which points to multiple columns. Each column associated with this row is a triple of: column name, column value and timestamp. For example, a personal database can capture different attributes for each person, e.g. name, phone, email, etc. and new attributes can be added to the database as they are provided with data. Typical use cases suited for column databases include data sets that can have multiple properties that are often not known ahead of time, e.g. keywords used to index web pages or text documents, tags dynamically assigned to user generated content, event logging. Sample implementations include: Cassandra, HBase, DynamoDB.

[22] Source: Wikipedia

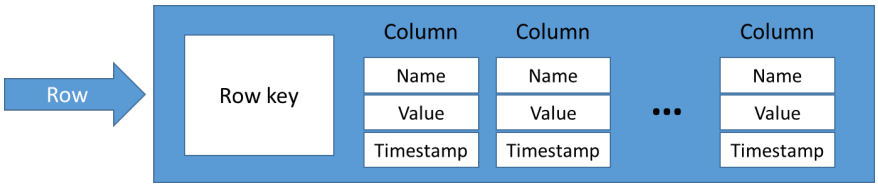


Figure 7: Row structure with variable number of columns

Graph databases store data as collections of nodes and relationships between them. They build on rich graph theory to efficiently query richly connected data. Nodes of a graph usually represent data entities that can have own properties (e.g. person entity with properties like name, email, etc.). A relationship connects two nodes. Relationships organise the graph and can have own properties as well. For example, two nodes representing people can be linked by different relationships, like friendship and employment. Typical use cases suited for graph databases include connected data (e.g. employees and their skills), recommendation engines (e.g. customers who searched this product also looked for those products). Example for clinical data could be a graph representing relationships between adverse drug reactions and involved substances and manufacturers. Sample implementations include: Neo4j, GraphDB.

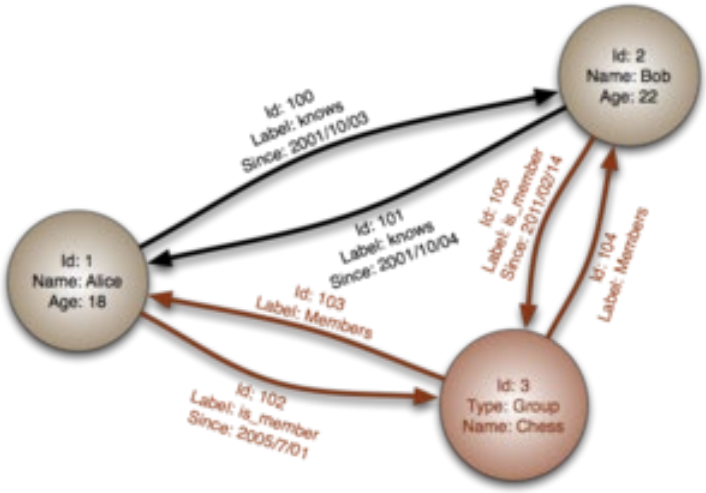


Figure 8: Graph database sample[23]

2.4.2. Hadoop ecosystem

Hadoop is another tool typically used in context of big data, sometimes even arguably used as a synonym for big data. In essence, Hadoop is an ecosystem of software packages that allows to store and process vast amounts of data in any format on large clusters of commodity hardware. NoSQL and Hadoop share some similar capabilities, but they are intended for different types of tasks.

In its original form Hadoop is just a distributed file system called HDSF for reliable and efficient storage of data files on multiple network nodes, i.e. machines across a computer network. These distributed files can be processed in parallel on multiple network nodes using the distributed computation

[23] Source: Wikipedia

mechanism called MapReduce. Because all storage and processing are performed on network nodes implemented on commodity computers, Hadoop can be easily scaled out just by adding new nodes.

Hadoop is very powerful for processing large datasets but requires programming skills and often significant effort to develop applications. Luckily, some higher-level tools have been developed to simplify processing of data and Hadoop maintenance. These products build upon or even replace HDFS and MapReduce to provide better performance, additional functionality (e.g. security), implement standard query languages and programming interfaces like SQL, JDBC, hide complexity of programmatically distributing data storage and processing across multiple nodes, etc.

2.4.3. Cloud big data storage

Hadoop gained popularity with the advent of big data. It allowed storage and processing of huge amounts of data in any format using commodity hardware. As a result, many organisations found themselves managing increasingly complex Hadoop infrastructure with multiple nodes and associated maintenance cost. The Hadoop software is predominantly free and commodity hardware cheap, but the experts managing this infrastructure are not.

This drawback of Hadoop is addressed by various cloud-computing providers offering big data solutions that do not require organisations to own any Hadoop infrastructure. These cloud providers implement various SQL and NoSQL databases that can be rented and charged based on usage time and actual usage of storage volume, processing power and data transfers. There is a huge and ever-growing number of cloud data storage offerings, so this paper mentions only few well-publicised examples:

Azure Data Lake Storage offered by Microsoft is optimized for storing massive amounts of unstructured data, such as text or binary data and it offers Hadoop compatible access.

Azure Cosmo DB is another cloud NoSQL database offered by Microsoft classified as multi-model database able to support document, key-value, wide-column, and graph databases.

Google Bigtable is a high-performance column-oriented data store that integrates easily with popular big data tools from Hadoop ecosystem.

Google Datastore is a NoSQL document-based database.

Google Cloud Spanner is a cloud database service built specifically to combine the benefits of relational database structure with non-relational horizontal scaling.

Amazon Aurora is a relational database built for the cloud that combines the performance and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases.

Amazon DynamoDB is a scalable NoSQL database that supports key-value and document data structures.

Amazon Neptune is a graph NoSQL database.

The above cloud offerings support both relational and various NoSQL data models. Selection of the right platform depends on requirements of a particular project, especially security requirements for data storage in the cloud and possibly network bandwidth and latency requirements for accessing large amounts of remote data.

2.4.4. Data integration technologies

Data coming from multiple diverse data sources needs to be integrated for analysis. This often requires transformation of data. Depending on the nature and format of the data source different tools and techniques can be used.

ETL is short for extract, transform, load which very well describes order of steps implemented to pull data out of one database and place it into another database. Extract is the process of reading data from a source. In this stage, the data is collected, often from multiple and different types of databases. Transform is the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. Load is the process of writing the data into the target database. Traditional ETL tools are very powerful and capable of working with relational data sources and many other structured or semi-structured data sources (e.g. web services, XML, JSON, CSV and Excel files).

ELT is a variant of ETL wherein the extracted data is first loaded into the target system and then transformed directly in the target database. ELT typically works well when the target system is powerful enough to handle transformations. Hadoop ecosystem comes with a whole set of tools that can be used for ELT and data capture, depending on the nature on the data source.

Data virtualisation platforms can provide access layer to diverse data sources. They can integrate multiple underlying data sources (e.g. relational database, data from web services and static Excel files, Hadoop or NoSQL database) into a single virtual data layer that can be queried and combined using standard tools, e.g. SQL or standard data analytical tools. Data virtualisation tools also allow implementation of additional access control and security mechanism, e.g. limiting access to particular data sets to certain users. This allows implementing security mechanism on top of data sources that do not support them natively (e.g. static CSV files or Hadoop data store). Another benefit of data virtualisation is that it allows creation of several virtual layers on top of the original data source, providing more data transformation and integration in each subsequent logical layer.

Service oriented architecture (SOA) and exposing data via web service that can be integrated by an enterprise service bus is another popular and mature integration mechanism.

2.4.5. Data warehouses and data lakes

A **data warehouse** is an aggregated storage designed for recurrent analytics/reporting on data typically originating from various data sources. Data warehouses implement predefined data models that follow either relational model known from SQL databases or a more specialised dimensional data model that aggregates data into facts and dimensions.

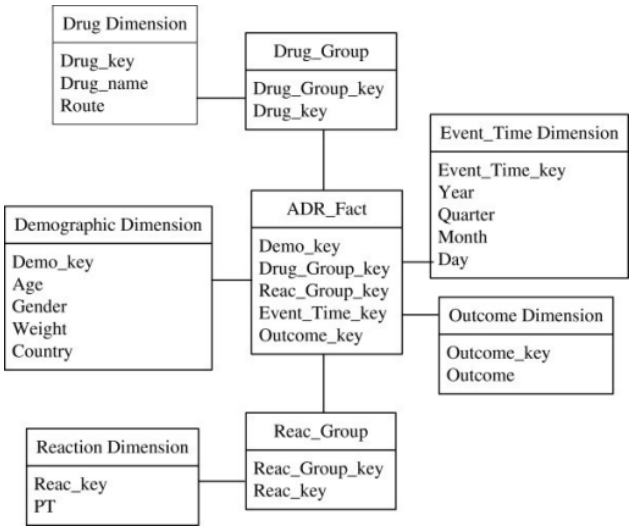


Figure 9: Sample dimensional data model

A data warehouse stores current and historical data in one single place. Typically, the data warehouse has an enterprise-wide scope, combining data from various business areas. Different communities and teams typically do not need the whole scope of the data warehouse for their analysis and extract only its relevant parts into data marts oriented towards their goals. They run their reporting on these specialised data marts. Each data source and the data warehouse and data marts implement their own data model.

Traditional data warehouses store in a single schema data integrated from multiple sources, both historical and present data. There are many data sources that do not easily fit into dimensional or relational schema, but which still need to be analysed and integrated. These data sources can have data in many formats, sometimes even unstructured data. These data sources can generate data at high velocity and volume (e.g. tweets, web logs).

Data lakes try to address the above problems. "A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. It's a great place for investigating, exploring, experimenting, and refining data, in addition to archiving data".[24] A data lake can also hold transformed data used for various tasks including reporting, visualization and analytics and machine learning. The data lake can store multitude of data formats: from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and even binary data (images, audio, video).

A data lake can become a database for all data used for analysis and reporting.

[24] James Serra, "What is a data lake?", accessed 4 Jun 2019

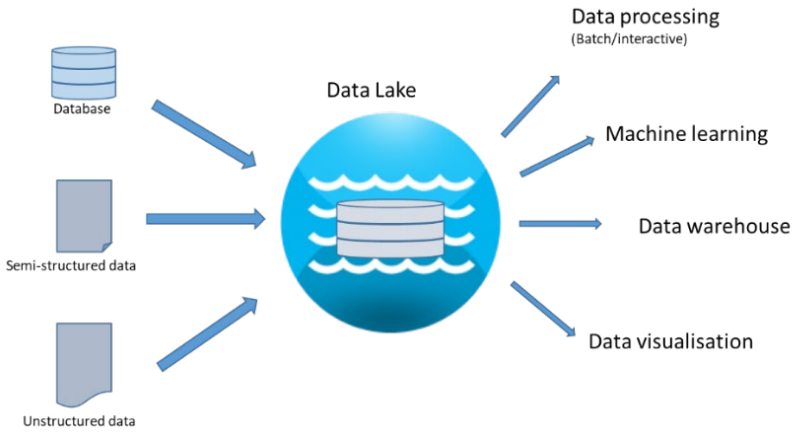


Figure 10: Data lake concept

Because data lakes can contain raw data in multiple formats it is necessary either to transform and cleanse the data for further analysis or to use analytical tools capable of working with such data. Any data scientist, developer or analyst working with data lakes must be aware of these constraints.

Storing all data in a single place can violate security policies of an organisation. Additional mechanisms may need to be implemented to limit access to data to entitled users only.

The ability to capture and store any kind of data in any format can quickly cause the data lake to fill in with low quality data from unknown sources. Also, there is temptation to capture all data, just in case, even if this data is not required for any analysis at the moment. Such approach can result in so called data swamp, a deteriorated data lake with little value to its end users.

Data lakes can be implemented using different platforms, incl. Hadoop and NoSQL, both on premise or in the cloud. These technologies allow to store a variety of data formats and scale to cope with increasing volume of data.

Main differences between a data warehouse and data lake can be summarised in this table:

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications.	Non-relational and relational from internet of things devices / sensors, web sites, mobile apps, social media, and corporate applications.
Schema	Designed prior to the DW implementation (schema-on-write).	Written at the time of analysis (schema-on-read).
Price/Performance	Fastest query results using higher cost storage in an underlying database management system.	Query results getting faster using low-cost storage distributed over commodity nodes.
Data Quality	Highly curated data that serves as the central version of the truth.	Any data that may or may not be curated (i.e. raw data).
Users	Business analysts.	Data scientists, Data developers, and Business analysts (using curated data).

Characteristics	Data Warehouse	Data Lake
Analytics	Batch reporting, BI and visualizations.	Machine Learning, Predictive analytics, data discovery and profiling.

2.4.6. Architecture

In a **data warehouse architecture**, data from multiple sources is extracted and transformed into an intermediate staging area, where data may be consolidated, cleansed. From staging area data is further transformed and loaded into a data warehouse. The data warehouse implements a consolidated data model that allows interpretation of data originating from diverse sources. It stores current and historical data in one single place. Typically, the data warehouse has an enterprise-wide scope, combining data from various business areas. Different communities and teams typically run their reporting on these specialised data marts. This architecture heavily depends on ETL based integration.

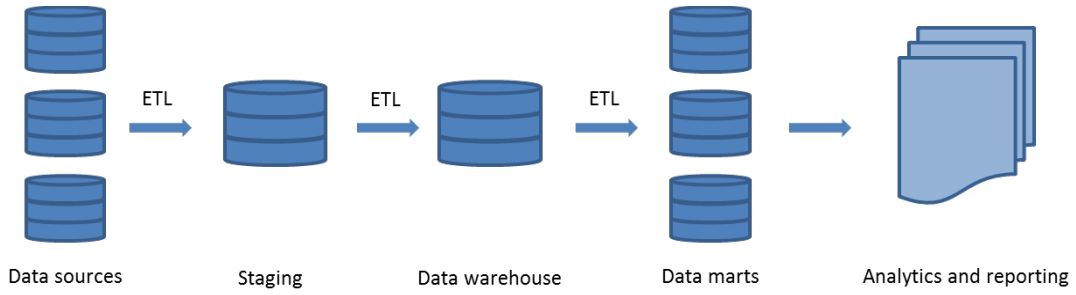


Figure 11: Classical data warehouse architecture

Traditional data warehouse reporting has been used for years and provides very good results. It is based on mature technology and has established best practices. It provides high quality of results achieved by consolidation and cleansing of data from multiple sources into a single picture in the data warehouse. This quality comes at a cost. It requires good understanding of data model of each data source, design of the consolidated data warehouse data model and implementation of business rules to transform from one to another. Typical project will include analysis and design phase, development and testing and finally deployment and maintenance. This takes time and involves IT development.

This architecture supports very well requirements for descriptive data analytics exposed in an analytics portal

A **data lake-based architecture** treats a data lake as low-cost, scalable environment to capture data in raw format from multiple sources and exposed for reporting, both traditional, static descriptive reporting for executive users and dashboards, and explorative data analysis done by data scientists. Data loaded into the data lake is transformed only minimally or not at all and the reporting layer must extract, transform and interpret data to provide meaningful insights. Raw data in the data lake is actively used for data experiments and exploration.

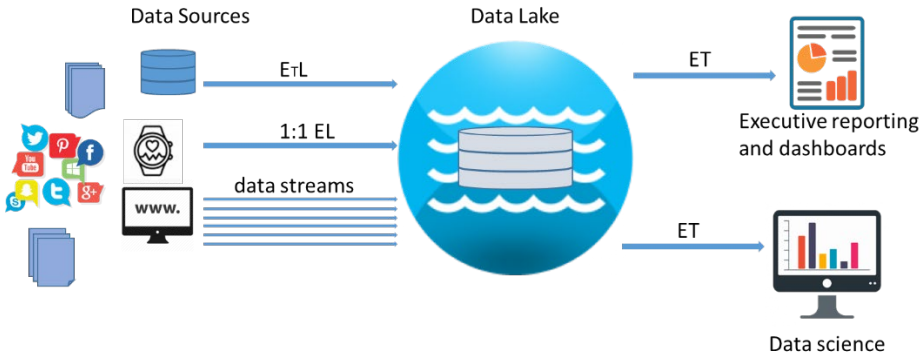


Figure 12: Simple architecture using data lake concept

In a **hybrid architecture**, a data lake can be integrated with a traditional data warehouse architecture described previously. In this case the data lake remains as the main source of raw data for data experimentation, but descriptive reporting is generated based on transformed and cleansed data in the enterprise data warehouse and data marts. Thus, the data lake becomes a staging area for the data warehouse.

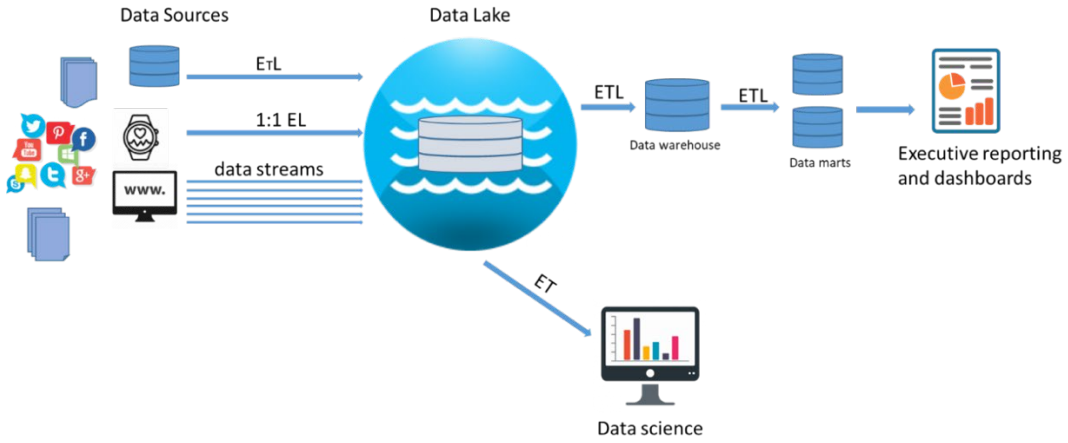


Figure 13: Hybrid architecture with both data lake and data warehouse

Different types and models of data analytics require an architecture that at the same time is able to support both traditional reporting as opposed to exploratory data science reporting.

A **bimodal architecture** defined by Gartner addresses requirement for an architecture that at the same time is able to support both traditional reporting and exploratory data science reporting. Mode 1 is a traditional IT, focused on stability and efficiency, while Mode 2 is an experimental, agile organization focused on time-to-market, rapid application evolution, and, in particular, tight alignment with business units. Experiments done in Mode 2 can often fail, but once they succeed and mature, typically they are moved into repeatable and production quality Mode 1.

Mode 1	Mode 2
• Traditional	• Fluid, flexible, agile
• Controlled	• Unforeseen
• Centralised	• Exploration Focus

Mode 1	Mode 2
• Production Scale • Architecture Focus • Send Exceptions to Mode 2 to simplify mode 1	• Failure is ok • Self – Service • Experiment • Move to mode 1 when standardised

Bi-modal data architecture for analytics can be implemented using **data virtualisation** technology described in Section 2.4.4.

This architecture uses data virtualisation layers to integrate existing data sources (like SQL databases, data warehouse and data files) and new data originating from NoSQL databases, web services and data lakes. Data warehouse and data lake can be maintained and used as described in Section 2.4.5. It is not revolution but evolution of exiting architectures that allows organisations to migrate to the bi-modal architecture.

Data virtualisation layer can have several virtual layers itself. The data source layer at the bottom is closest to the original data sources and exposes data in their native format, without any transformations. Its main role is providing uniform access control and management of data sources that natively do not support this functionality (e.g. flat files or a Hadoop based data lake). Enterprise data layer exposes integrated and cleansed data, providing uniform enterprise view of underlying data sources. The top data consumption layer presents specialised and highly curated data view prepared for particular use cases and business domains (in that respect it is a virtual counterpart of business specific data marts described in Section 2.4.5.).

Traditional, descriptive reports exposed in the analytics portal mostly will be based on data exposed through data consumption layer. Self-service users of the analytics workbench will be able to access data both from the data consumption layer and enterprise data layer. Since data exposed by these layers has already been pre-processed in the data virtualisation layer, it is expected to be of good quality and managed. Finally, data scientists using the data lab will be able to access and combine raw data from the data lake and other sources to run their simulations and analytical models.

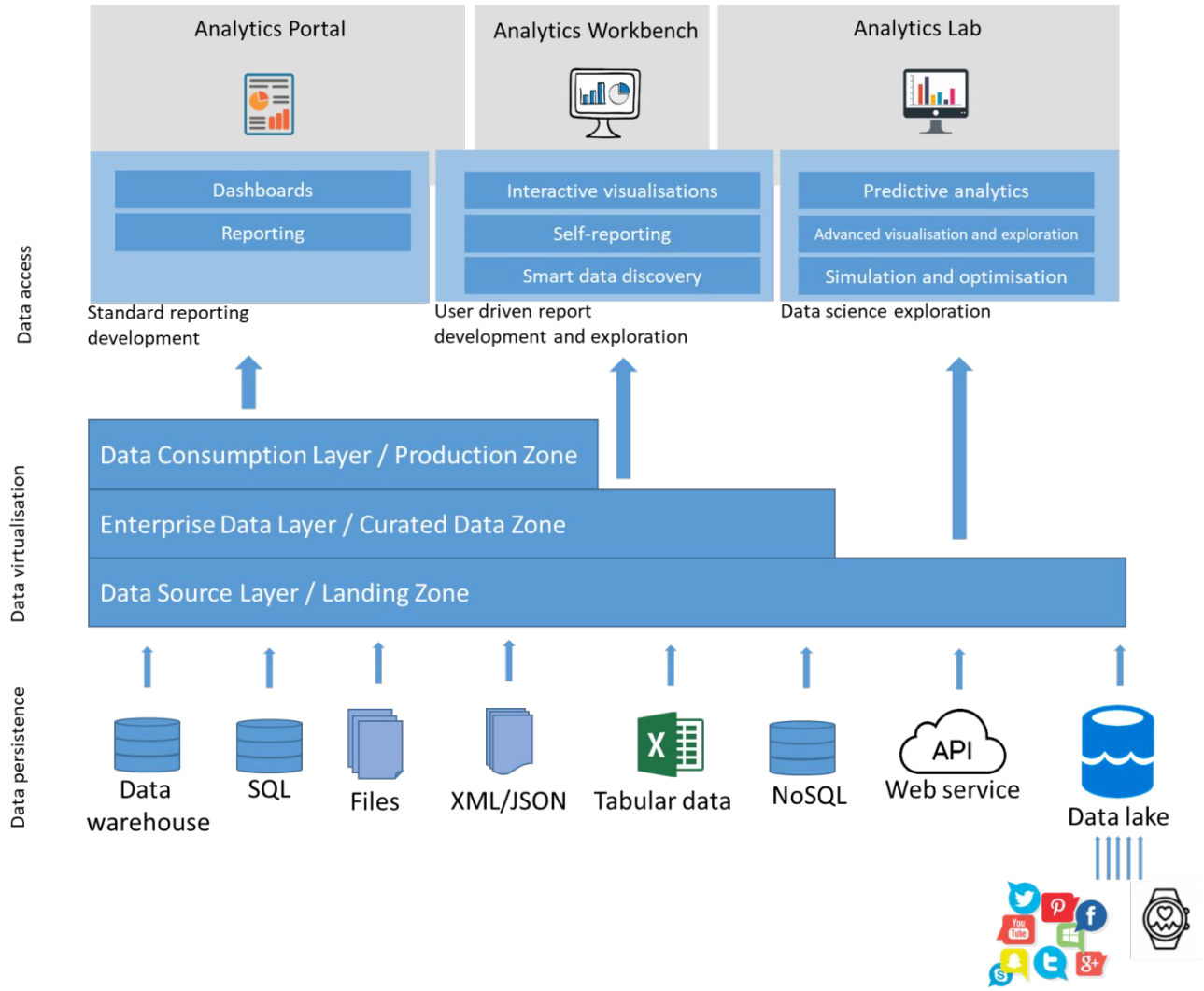


Figure 14: Bi-modal data analytics architecture using data virtualisation

2.4.7. Related concepts and technologies

There are other IT concepts and technologies that make the move to big data analytics easier.

Computer virtualisation is a concept that enables better use of physical IT infrastructure and shortens time necessary to deploy and configure computer environment. In its basic form virtualisation allows to programmatically simulate a whole computer as software running on another computer. Typically, several virtual machines can run in parallel on a single physical machine and share its resources like processor, memory and disks space. From the end user perspective, a virtual machine behaves just like a normal computer – it will have own operating systems (e.g. Windows or Linux) and will allow to install and run software just like on a physical machine. Because multiple virtual machines can work on parallel on the same physical computer, it typically offers better utilisation of its hardware. It is possible to pre-configure a virtual machine template with preinstalled software (e.g. Hadoop) and use it to create multiple similar virtual machines. This shortens the time to have a preconfigured machine running, e.g. to dynamically add new computing nodes to a Hadoop cluster. It is possible to run virtual machines on a desktop or even a laptop computer. For more resource intensive tasks, virtual machines will typically run on a dedicated strong server which can be located in organisation's server room or rented from an external provider.

Virtual machines running on external infrastructure are an example of cloud computing where basic infrastructure (like processor, memory, hard disk) is provided to many customers as a service. Infrastructure as a Service (IaaS) cloud providers allow to run multiple virtual machines on their infrastructure and create new ones on demand.

Cloud computing allows using storage and processing power offered over internet by external service providers. Cloud computing's benefit is that it can be acquired quickly and dynamically sized to meet current requirements. Cloud computing typically is priced based on actual usage and as such is an attractive alternative to procurement of own infrastructure.

Cloud computing comes in several flavours:

Infrastructure as a Service (IaaS) offers basic infrastructure, typically virtualised machines where users need to install and configure own software and services. This typically requires IT involvement in managing the operating system, software patches, etc.

Platform as a Service (PaaS) provides computing platforms which typically include programming language execution environment, database and web server. Users can use this infrastructure to deploy own applications. Typically, there is no need to worry about underlying virtual machine and operating system, reducing infrastructure maintenance effort.

Software as a Service (SaaS) offers preconfigured software that can be directly used and replaces traditional on-device software.

Various vendors, trying to differentiate themselves in the market are offering Big Data as a Service or Data Analytics as a Service, but these typically are variations of PaaS or SaaS with preinstalled and pre-configured Hadoop or other big data technologies.

Metadata is data about data, typically business content data. It contains all information describing how and where data assets are structured and stored, and how they are transformed and moved between various systems. Metadata management is the end-to-end process and governance framework for creating, controlling, enhancing, attributing, defining and managing a metadata. **Metadata management tools** support this process by providing mechanisms to discover, describe and link data assets in an organisation. Typically, these tools will provide data lineage information from end user report back to the original data sources.

Self-service data analytics tools are a form of business intelligence tools that can be used by business power users to create queries and generate reports without support from IT or data science teams. These tools are often characterised by simple to use interface (often web based), with basic analytical and visualisation capabilities. These tools often work with simplified or pre-processed data models scaled down for ease of use and access. Many organisations combine self-service data analytical platforms with data virtualisation or metadata management tools to increase discoverability, understanding and accessibility of data sets exposed to business power users.

Blockchain is another technology rapidly gaining popularity. It relies on strong cryptography[25] and decentralised, publicly accessible digital ledger that contains tracks of all approved transactions. When one of the participants wants to make a transaction and add it to the ledger, he/she calculates and signs a small block of digital information describing this transaction. Then the participant broadcasts this block to the network. Those in the network compare the transaction with the current state of the

[25] Cryptography is a method to protect data and includes both encryption (which is reversible) and hashing (which is not reversible, or "one way"). Strong cryptography is secreted and encrypted communication that is well-protected against cryptographic analysis and decryption to ensure it is readable only to intended parties

ledger and if correct approve it. Once the transaction is approved, the block representing it is added to a chain of previous transaction blocks that constitute the ledger (thus name – blockchain).

All blocks in the chain are connected using strong cryptographic algorithms. It is not possible to change one block without influencing other blocks. The blockchain representing the ledger is distributed in the network making it difficult for one party to falsify the whole chain. There always will be other copies proving the forger wrong. Thus, data represented by the blockchain can be considered trustworthy. On the other hand, managing blockchains due to use of cryptographic algorithms can be computationally intensive, what may limit its usability for large data sets.

It is also possible to connect external data to the blockchain. In this case a digital fingerprint (hash) of this data is inserted into the chain. The veracity of the external data can now be controlled – if it no longer matches the hash it has been tampered with. Using this method, it is possible to secure sensitive data (e.g. data from individual patients in a trial) without actually having access to it. Blockchain technology can be used in scenarios where trusted data needs to be readily available and there is not possible or desirable to have a central authority storying this data. An example may include information about supply chains to prevent distribution of falsified medicines.

Internet of Things (IoT) refers to the increasing number of network-connected devices, from simple sensors to smartphones and wearables, that can send and receive data. These devices can provide high volume and velocity of data, often real-time data that is produced in a continuous stream. Capturing, storying and processing such data streams are challenges many of the described big data technologies tried to address. Connected devices and sensors have a huge potential for health care as they can provide valuable data on patients and their environment, often called Patient-Generated Health Data.

Edge computing refers to storying and processing data locally, on a network edge, rather than transmitting to a centralised data centres. It keeps data and processing closer to the user and allows to minimise latency of transmitting potentially huge amounts of data. Edge computing is often mentioned in context of Interment of Things, where data captured by the sensors is processed and analysed locally. Edge computing and IoT have potential for healthcare, for example allowing to provide rural medical help, faster response to sudden changes in a patient's condition. The number of potential use cases is growing[26].

2.5. Opportunities

There are many examples where big data analytics can provide benefits for healthcare. For example it can help detect early epidemic based on analysis of web searches for symptoms; analyse electronic health records to reduce duplicate tests and improve patient care; analyse trends in hospital care; support evidence-based medicine to match symptoms to a larger patient database in order to come to an accurate diagnosis faster and more efficiently. Other examples are countless and depend only on availability of data and ingenuity of a person analysing it.

Information technology can provide technical solutions, architectures and best practices to help data scientist and business users access and analyse big data in a controlled manner. These technical solutions must be supported by proper management processes and data governance.

[26] See: https://www.vxchnge.com/blog/edge-computing-use-cases-healthcare, accessed 4 Jul 2019

2.6. Challenges

The volume, variety and veracity of big data are the biggest challenges for its analysis and value extraction. The sheer amount of data makes it like looking for a needle in a stack of hay. Different data sets often need to be combined and compared what requires some kind of normalisation to avoid comparing pears and apples. In many cases decision based on analysis of big data must be justified, i.e. lineage of data used for analysis must be known and all transformation and analysis steps must be clearly documented. Collecting any and all data into a single place like data lake in an uncontrolled manner quickly turns it into unmanageable data swamp.

All these problems are highlighted by results of a survey[27] revealing that data scientists spend around 80% of their time on preparing and managing data for analysis. This includes 19% of their time dedicated to identifying and collect input data sets and 60% of their time on cleaning and organizing data.

In many organisations there is never enough data scientists and as described above, they sacrifice majority of their time for tasks other than actual data analysis. This leaves many relatively simple data analytics and reporting tasks unsupported by data scientists and forces technology savvy business power users engage in building diagnostic or predictive analytics typically by using tools like Excel. These power users typically know their own business data but have very limited access to other data sets and do not have proper tools for importing, cleansing and merging data.

Security is another important challenge for processing big data. Access to commercially sensitive data must be controlled. European legislation like GDPR puts further controls on how personally sensitive data must be processed.

At a technical level many organisations have troubles in building and maintaining own infrastructure capable of efficiently capturing, storying and processing high velocity, volume and variate data. Building own big data infrastructure, e.g. using Hadoop, requires skills, time and money.

[27] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/, referenced 29 March 2019

2.7. Recommendations

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
1	Governance	Establish a working methodology to leverage analytics opportunities	Create a data analytics team that incorporates multiple profiles (e.g. data scientists, data architects, data stewards) to advise business units on how best to process and analyse their data to add value. Expose non-sensitive open source regulatory data sets to provide opportunities for researchers to develop novel and possibly disruptive analytical approaches. Establish best practice to implement data governance in parallel to introduction of new big data technologies to ensure data quality, data lineage and traceability of analysis results.	Creation of a data analytics team Include in IT architecture documentation aspect on data governance including skills needed and possibilities of data sharing
2	Technology	Implement an architecture that facilitates data exploration and experimentation	Implement Bi-modal architecture to provide a space for data exploration and experimentation and turn successful experiments to repeatable and production quality models. Implement a self-service data analytics IT framework to encourage business power users to perform queries and generate reports on their own with nominal IT or data scientist support. Implement data virtualisation technology to be used to increase data discoverability and reusability of cleansed data sets. Embrace cloud technology for building big data and analytics infrastructure. Provide for adequate data transfer speed or minimise movement of data.	Increased use of the technology relevant to support Big Data IT architecture

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
			Consideration for data security and privacy must precede any adoption of cloud technology. Implement security by design. Implement metadata management tools to support implementation of data governance processes.	

Data Analytics – Data manipulation

3. Data manipulation

3.1. Why

Data manipulation is a fundamental step in data analytics; it is often said that 80% of a data analyst's time is spent on making the data ready for the analysis, particularly so where data is unstructured, such as social media data or any free text data, but also in structured data.

The way data is aggregated or recoded, for instance, can have an influence on the results. These can stem from unintentional mistakes, or be intentional such as *p*-hacking, data fishing or data dredging, which typically involves some sort of data manipulation, such as grouping or stratifying to make a *p*-value either significant or non-significant (i.e. making use of Simpson's paradox). While this occurs at data analysis, transparency in data manipulation reduces the likelihood of such mistakes.

With increasing sizes of databases, increasing abilities to link data sets and with novel opportunities to collect unstructured data, such as social media data, it is important to consider best approaches in data manipulation to ensure efficiency, scalability, reproducibility and transparency.

An example of the importance of data manipulation and an illustration of its relationship with data standardisation and analytics, can be seen with the Observational Medical Outcomes Partnership's Common Data Model initiative.

The experienced analyst will be familiar with literature on data manipulation focused on the use of a specific statistical coding language such as also SAS®, R, Python, among others. This chapter is purposefully coding language-agnostic; however, the reader will quickly realise that the suggestions made fit naturally within the open source language paradigm.

3.2. Objectives

This chapter provides a brief explanation of where data manipulation sits within an analysis pathway and how it can permeate across the different stages of the analytical pipeline.

The high-level nature of this topic assumes the reader is a junior analyst, scientist or assessor looking for guidance, reference terminology and a list of useful literature. However, this text may be equally useful for a seasoned analyst looking for EU network-wide practical suggestions that can serve as consensus within the regulatory network.

One primary concern of this document is to ensure some standardisation of terminologies for regulatory purposes.

It also describes common data manipulation processes and presents options on how best to implement them in an analytical pipeline such that they are regulatory sound, reproducible and transparent.

An appropriate balance between the technological or computer science perspective and the analytical or data science perspective is attempted throughout the text.

3.3. Main concepts

Data manipulation can be defined as the act of transformation of *raw* data aimed at allowing patterns in the data to emerge.

Terms such as data wrangling, data munging, data processing and data preparation are often used interchangeably with data manipulation.

For the purpose of this document, the term data manipulation is preferred to refer to all data transformation operations performed on *raw* data with the aim of making it research ready.

These activities include acquiring data, updating, adding, removing, sorting, merging, shifting, aggregating and often, some low-level analytical work, undertaken to understand the data.

Noticeably, low-level analytical work to understand the data may lead to meta-data which can be considered as raw data for another analytical pipeline.

Formally, data manipulation is an intermediate step in the analytical framework however practically, because the analytical pipeline is often data-driven, stepwise, recursive and/or iterative, data manipulation is performed across the analytical pipeline.

3.4. Glossary

Data aggregation – actions that reduce and rearrange data with a view to summarising it, e.g. grouping raw data into yearly counts.

Data manipulation – act of transformation of *raw* data aimed at allowing patterns in the data to emerge; includes acquiring data, updating, adding, removing, sorting, merging, shifting, aggregating and often, some low-level analytical work, undertaken to understand the data.

Data munging (see data manipulation).

Data preparation (see data manipulation).

Data processing (see data manipulation).

Data transformation - Data transformation is a rare example of a term that can be interpreted differently depending on whether the reader's background is in statistics/data science or in computing. For the purpose of this text, the statistical application of the term will be used, and data transformation will be considered the application of a mathematical function to all values of a variable for making it suitable for an analytical process, e.g. taking the logarithm.

Data wrangling (see data manipulation).

Incorrect data – incorrect data are erroneous values assigned to variables in observations. Erroneous values could be specific to a variable's range or category list, e.g. an age of 800 is incorrect data. It could also reflect incompatible values, which are within the correct range or category list of different variables, e.g. male patient recorded as pregnant.

Metadata – data that provides information about other data, i.e. metadata provides an understanding of raw data which helps unlock its value.

Missing data – data values that are not stored for a variable in the observation of interest. There are several types of missing data:

- Missing completely at random (MCAR) the probability that the data are missing is not correlated with any other variable in the set of observed variables.
- Missing at random (MAR) is where the missing data can be predicted based on other observed variables but not the outcome variable. This means that an advanced imputation method, such as multiple imputation, could be used to fill the null value with a probabilistic value.

- Missing not at random (non-ignorable) is where the outcome variable is correlated with the missingness.

Raw data – unanalysed data, collected from a source. Meta-data or the output of one analysis can be considered raw data for another analysis.

Rectangular data format – a data format of a flat file composed of columns (variables) and rows (observations).

Variable – Any quantity that varies. Any attribute, phenomenon, or event that can have different values. A column in rectangular data formats. Variables that are not the outcome of interest in an analysis are also known as predictors or as features.

3.5. Overview

The extent of data manipulation in a given analysis may vary: data collected from structured data models or from clinical trials may require fewer transformations as compared to real-world data, in particular those stemming from novel data collection opportunities, such as social media data, where advanced data manipulation techniques may be required.

3.5.1. Data types

Data type is a data storage format that can contain a specific type or range of values. Different programming languages and statistical coding languages have the same or similar types but might name them differently. In general, data types can be split into primitive (or built-in/basic) types, non-primitive (or derived/composite) types and abstract types.

Table 1: Data types

Data types	Examples
Primitive data types Primitive data types are the most basic data storage formats.	**Boolean** is data type that only stores two values, typically true and false, and is meant to represent truth-values. **Integers** are any whole number, which can be positive, negative, or zero and which do not have decimal places. Integers result typically from aggregated count data, such as the count of patients that met a pre-defined endpoint in a clinical trial. **Real values**, sometimes defined as floating point numbers, as they are numbers that contain a floating decimal point are typically fractions of integers or measurements in a continuous scale, such as height and weight. **Characters** are any letter, number, space, punctuation mark or symbol that can be typed on a computer. The list of characters that can be typed (or interpreted) is defined by the ASCII and extended ASCII set. **Date/time** are special integers that reflect data and time.

Data types	Examples
Non-primitive data types Composite data types are data types constructed from primitive data types.	**Arrays** are any set of data that is grouped together using the same identifier. Typically, these elements are all of the same data type, such as an integer or string. **Strings** are a unidimensional array comprised of a set of characters and may include whitespace (i.e. spaces). **Two-dimensional arrays** are tables of data that use two levels of indexing an array. **Lists** are similar to arrays and are sequences of ordered data types.
Abstract data types Abstract data types are any types of data that do not specify an implementation.	Examples include hierarchical structures, such as trees, graphs, among others.

Different statistical programming languages (e.g. R, Python) interpret these data types and have data type specific functions and methods. Hence, understanding the data types helps understand what functions and methods can be performed.

For instance, adding, subtracting or multiplying integers always results in an integer, however dividing two integers may result in a different type of data – real values. Furthermore, date data types, while integers, cannot be averaged.

Most often statistical programming languages will try to determine the type of data based on the data uploaded. For instance, a table uploaded with two columns, one with street names and another with a solely numerical postcode is easily interpretable for a human as a string and a character data type. When sourcing the data, if the type of data is not specified, a software may assign the numerical postcode as an integer or maybe as time data type.

Noticeably, the data type of a variable can often be determined by profiling all the data values in the column. Some advanced tools can discover patterns and outliers (e.g. 99% of values follow the format 99-999, where 9 stands for a number 0-9, but 1% has a different pattern, such as a letter instead of a number, 12-34P).

These types of data manipulation errors are obvious and may even seem trivial, and mostly do not require a trained eye to be detected. However, these are common errors, which may go unnoticed particularly in large data sets and those that are imported across software systems.

Understanding the data types is thus fundamental and an important task in data manipulation. Most software will typically do a good job at assigning data types and the errors can be corrected reactively, once the error is detected. However, the use of big data and multiple software in the same analytical pipeline may lead to undetected errors.

Consider a spreadsheet that has a set of relevant codes one wants to use, e.g. 7M0h300 and 1517.00. Using a spreadsheet software that assigns data types to individual cells, and not to the whole column, would mean the code 7M0h300 will be treated differently to 1517.00: the first will be considered a character whereas the second might be considered a integer and converted to "1517", i.e. ".00" will be rounded.

Importing these two codes from the spreadsheet to another analytical software will turn the code 1517.00 into 1517. Thus, if one tries to, for instance, to filter a data set with these codes, 1517 will not match with 1517.00.

While this problem seems negligible with two codes only, it is not uncommon to use hundreds of codes in an analysis, which makes the task of identifying possible pitfalls before running an analysis all the more relevant.

Hence, actively checking and assigning data types – in some software also called class – is recommended: 1) after data collection or 2) after importing from other formats. This is particularly true for files that do not store self-describing data such as comma separated value files.

3.5.2. Reshaping Data

Data reshaping is the action of rearranging data into a different, more convenient format for analytical purposes. It may be helpful to understand data reshaping as analogous to ***data aggregation***.

Aggregation is a common task where the data is both reduced and rearranged. Extracting monthly counts of events from a database of patient outcomes is one example of aggregation. Reshaping equally involves rearrangement but reshaping should preserve the original information. Common data reshaping tasks include splitting, merging or joining and changing rows to columns.

Data often has multiple categories in a variable. Analysing their relationship, through for instance a regression model, typically requires reshaping the raw data.

This is because categorical variables, contrary to dichotomous or continuous variables, cannot be applied into the regression equation in their original form. Each categorical variable needs to be recoded into a series of variables that can then be entered into the regression model.

This format is not ready for a regression:

subject_id	**Gender** **Levels (2):** **Male (M)** **Female (F)**	**smoking_status** **Levels (3):** **non-smoker** **past smoker** **smoker**
1	M	Past smoker
2	M	Smoker
3	F	Non-smoker
4	F	Smoker
5	F	Non-smoker
6	F	Non-smoker

subject_id	Gender Levels (2): Male (M) Female (F)	smoking_status Levels (3): non-smoker past smoker smoker
7	M	Past smoker
8	M	Past smoker

This format can be applied to a regression:

subject_id	gender_male	past_smoker	smoker
1	1	1	0
2	1	0	1
3	0	0	0
4	0	0	1
5	0	0	0
6	0	0	0
7	1	1	0
8	1	1	0

The format not ready for a regression is sometimes called the ***long format*** (or narrow, stacked or tall), where multiple variables are in a single column as exemplified above. Conversely, the format ready for a regression is sometimes called the ***wide format***. Note that certain software packages (or libraries or macros) may include code to transform a table from long format to wide format, and thus the analyst may not realise that recoding has happened.

Reshaping the data usually comes with a process of recoding. The one demonstrated above is named ***dummy coding***. This recoding takes a categorical variable with *n* possible values and recodes it to *n*-1 columns with 0 and 1, based on the presence (1) or absence (0) of the value in the observation. The reference level will be coded as 0, in this case Female is "*gender_male* = 0" and *non-smoker* is "*past_smoker* = 0" AND "*smoker* = 0".

There are other forms of recoding particularly for regression purposes that are not discussed in this section.

At this stage, it is important to consider the type of database management system, particularly its classification based on data model. There are several database management systems, better described elsewhere, namely relational, examples of which include Oracle®, Microsoft SQL Server®, IBM BD2®, MySQL®, SQLite® and PostgreSQL®, hierarchical such as an eXtended Markup Language document (XML), network systems, object-oriented such as IBM DB4o® and flat file-based systems.

In most circumstances, reshaping the data will also change the model of the data from a normalised relational model, where data are organised in a set of formally defined tables linked to each other by some identifier, to a flat table. Data from several tables extracted from a relational database is typically reshaped into one wide table, often called ***analytic base table*** that can be further reshaped or analysed using, e.g. regression models.

The data reshaping steps will depend on the type of data and on the purpose of the analysis. Splitting the data is for instance, less useful in a descriptive study that in predictive analytics.

A basic requirement of machine learning methods, for instance, is that the data is randomly split into training and testing samples, at an 80:20 split for instance. Models are then trained on the training dataset and then their accuracy, or other metric of acceptance, is assessed on the testing dataset. This split reduces the chance of overfitting the model that would then affect accuracy in real world implementation.

Data permitting, more elaborate methods of splitting data described in the machine learning chapter, may be employed. In these cases, the training dataset is further split into training and validation datasets, which after training allow for fine-tuning the model, prior to testing in a fresh set of data.

This transformation exercise as part of a machine learning method illustrates once more the overarching aspect of data manipulation across the analytical pipeline.

Some readers will be familiar with the pivot table in spreadsheet software, which provides a very intuitive way of reshaping data and could empirically be considered the most frequent tool for reshaping. However, graphical user interfaces may result in untraceable errors in the analytical pipeline – there is no auditable data manipulation history – and may affect result, interpretation and reproducibility.

In summary, **the reshaping steps should preferably be auditable**, particularly where it concerns big data. It then follows that the analytical database table(s) created from reshaping should be saved as a separate data set.

Naming conventions

While naming conventions are best placed in style guides, with data reshaping, new variables are created and thus one should be mindful of naming conventions.

There are several ways to create a variable name:

Table 2: Naming conventions

Type of naming conventions	Application	Example
camelCase	Whitespaces between words are removed and the first letter of the following word is capitalised.	startDate

Type of naming conventions	Application	Example
PascalCase	Same as camelCase but where the first word is also capitalised.	StartDate
snake_case	Whitespaces between words are replaced with underscores.	start_date
kebab-case	Whitespaces between words are replaced with dash.	start-date
Dot notation	Whitespaces between words are replaced with periods.	start.date

3.5.3. Transforming Data

Data transformation is a rare example of a term that can be interpreted differently depending on whether the reader's background is in statistics/data science or in computer science. For the purpose of this chapter, the statistical application of the term will be used, and data transformation will be considered the application of a mathematical function to all values of a variable for the purpose of making it suitable for an analytical process. Thus, taking the logarithm, square root, reciprocal, or applying some other function on the values of the data is data transformation.

Computer engineers will talk of data transformation to mean data reshaping and use other terms, such as data pre-processing, to mean statistical transformation. For example, changing from a long format to a wide format, i.e. reshaping data, is considered a transformation of data.

There are several reasons why one would perform data transformations, which can be reasonably grouped in two main reasons: to perform some analytical technique that requires data to follow a distribution of a particular kind, such as the Gaussian distribution, and the variance of different groups of observations to be uniform; or to present data in meaningful graphical representations.

Assume a plot of cases over time for two safety outcomes, skin rash and drug reaction with eosinophilia and systemic symptoms (DRESS), is performed to assess a certain intervention and the relationship between the two outcomes. If rash has thousands of cases per unit of time and DRESS has only a handful of cases per unit of time, the variability of cases will seem more evident for the rash cases, whereas the DRESS cases, due to the y-axis scale will appear roughly constant. In this case, using a logarithmic scale might improve data visualisation.

It should be noted that parametric tests are robust to some deviation from their assumptions. Data does not have to be perfectly normal and homoscedastic (i.e. the variance of the error terms, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable, is constant and does not depend on the values of the independent variables). Some basic data transformation methods are presented below (**Table 3**).

Table 3: Data transformation examples

Method	Function	Applicable in	Limitations
Log	$\ln(x)$	Right skewed data (positive skew) Improving visualising variability	Zero values Negative values

Method	Function	Applicable in	Limitations
	$\log_2(x)$ $\log_{10}(x)$		
Log	$\ln(x + 1)$ $\log_2(x + 1)$ $\log_{10}(x + 1)$	Right skewed data (positive skew) with zero values Improving visualising variability	Negative values
Square root	$\sqrt{x}$	Right skewed data Positive values	Negative values
Square	x^2	Left skewed data	Negative values
Reciprocal	$1/x$	Making small values bigger and big values smaller	Zero values Negative values
Z-score standardisation or standard score	$(x-\mu)/\sigma$	Putting data from different sources onto the same scale	Data with many outliers
Scaling normalization or feature sclaing	$(x_i-X_{min})/(X_{max}-X_{min})$	Putting data from different sources onto the same scale	Data with many outliers
Differencing	y_t-y_{t-1} where *t* is time	Transforming a non-stationary process to a stationary process	Not possible to assess change in mean

Certain transformations are iterative. Principal component analysis, for instance, which is better described in the machine learning section, is an orthogonal transformation that is useful in reducing the dimension of data (e.g. the number of variables or features), and typically requires raw data to be normalised before being performed.

Transforming outcomes and/or predictor variables has the potential to circumvent a number of problems with models. In practice, data transformation may require a trial-and-error approach. A transformation is applied, and the model is tested, and this is carried on cyclically until problems with the model are eliminated.

3.5.4. Dealing with missing data

Missing data is one of the most common problems in data analytics. Null values are immune to data transformations and are not taken into consideration for the majority of statistical models and thus are non-informative as regards to the hypothesis being tested.

This does not mean that missing values are not useful. Missing data may be:

Missing completely at random (MCAR) the probability that the data are missing is not correlated with any other variable in the set of observed variables.

Missing at random (MAR) is where the missing data can be predicted based on other observed variables but not the outcome variable. This means that an advanced imputation method, such as multiple imputation, could be used to fill the null value with a probabilistic value.

Missing not at random (MNAR) or non-ignorable is where the outcome variable is correlated with the missingness.

Non-random processes in missing data may provide valuable insights that otherwise would not be identified, it could show, for instance that in an electronic health registry there are proportionally more missing data on smoking status in younger patients without cardiac history because medical doctors are more concerned about smoking status in elderly or patients with cardiopathies.

Thus, prior to tackling the missing values in the statistical models chosen, it is useful to understand the process underlying the missingness, particularly if a high proportion of missing values is present in the main predictor variables, by for instance, considering the probability distribution of missingness, i.e. if certain sub-populations have higher likelihood of having missing values.

There are roughly three general approaches to handle missing data:

Recovering the missing values: where the participants/reporters/investigators are contacted and asked to fill out the missing values. This is what happens for instance in follow-ups of individual case safety reports of adverse drug reactions.

Performing listwise or pairwise deletion: deleting data from any observation with missing values, particularly where the sample is sufficiently large that this data can be dropped without substantial loss of statistical power. However, care must be taken not to inadvertently remove a class of participants due to not random missingness.

Imputation: is where a value is assigned in replacement of the missing value, according to some rule or algorithm. These include educated guesses, average imputation (i.e. using the mean), common-point imputation (i.e. using the middle point or median), regression substitution and multiple imputations (similar to regression substitution but using multiple variables and taking advantage of correlations between the responses). This list is not exhaustive and there are other sophisticated methods, such as data augmentation, best described elsewhere (see resources section).

These approaches are dependent on the type of missingness. As one would expect, imputation for variables missing completely at random seems less useful than using imputation in variables missing at random, where some correlation with observed data is present.

Type of missingness	Approaches (in suggested sequence)	Comments
MCAR	Recover missing values Listwise or pairwise deletion	In variables with MCAR, imputation is not useful as it would be a purely random process.
MAR	Recover missing values Imputation Listwise or pairwise deletion	In the context of medicine regulation, a case could be made for using imputation in MAR variables however, it may not be entirely clear if the variable is truly missing at random and further uncertainty may be introduced with imputation.

Type of missingness	Approaches (in suggested sequence)	Comments
MNAR	Recover missing values	MNAR variables are informative as the absence of values is related to the outcome of interest. Using deletion methods will introduce a bias. It is better to recover the missing values or to understand the relationship with the outcome.

Similarly, to understand what type of missingness is present, one needs to diagnose the mechanism of missingness. There are several ways to reach a degree of understanding of the mechanism behind missingness, which are beyond the scope of this section however, at a minimum graphical methods could be employed.

Some data standards, such as HL7, allow the provision of additional data on why data is missing, so called null-flavour, e.g.:

NI - No information. This is the most general and default null flavour.

NA - Not applicable. Known to have no proper value (e.g., last menstrual period for a male).

UNK - Unknown. A proper value is applicable but is not known.

ASKU - Asked, but not known. Information was sought, but not found (e.g., the patient was asked but did not know).

NAV - Temporarily unavailable. The information is not available but is expected to be available later.

NASK - Not asked. The patient was not asked.

MSK - Masked. There is information on this item available, but it has not been provided by the sender due to security, privacy, or other reasons. There may be an alternate mechanism for gaining access to this information.

OTH - Other. The actual value is not an element in the value domain of a variable. (e.g., concept not provided by required code system).

3.5.5. Dealing with incorrect data

Equally, it is possible that manifestly incorrect entries are introduced in the data, for instance an erroneous value assigned to a variable such as "*age* = 800 years", or incongruent or incompatible information in more than one variable, such as the occurrence of a medical event prior to birthdate or conception.

Handling such cases may be tricky as the value for one variable may be wrong but the remainder of the variables could be correct and informative. In such cases, several steps similar to how to deal with missing values can be performed:

Do nothing and accept the error as a limitation in an analytical process that uses this data

Flag incorrect values (i.e. by creating an additional column or meta data) so that the analysis can then be done either keeping the error or excluding the case with incorrect values (listwise deletion).

Nullify incorrect values.

Impute a substitute entry.

Restore or try to correct the data.

3.5.6. Metadata

Metadata is defined as data that describes other data. Metadata is present in virtually all sources of data. Authors, dates and times, formats, titles, etc., are metadata common to Portable Display Format (pdf) documents, Word documents, emails, published articles and photos. Conversely, in a database, typical metadata include the description of the columns or variables present, such as data type, length, description, levels, relationships, etc.

Metadata falls into three main categories:

Descriptive, used for discovery and identification and including information such as title, author, abstract and keywords.

Structural, which describes how information is put together – page order to chapters, relationships between digital content, etc.

Administrative, which provides information for better resource management, such as when and how the resource was created.

There is value to be extracted from metadata, primarily from being able to better unlock value from the data itself, by understanding the structure and relationships between the data and metadata. In fact, anecdotally metadata is considered to be "little data" that help individuals understand big data better.

A simple example of where metadata can assist is the sharing and merging of data sets. For the best possible use of the data, the receiver requires understanding of how the data was collected, put together and what each variable means. In fact, robust metadata documentation is an often proposed alternative to data transformation into common data models. Particularly where that standardizing may lead to some loss of information and thus reduce the ability to fully leverage the data in a dataset.

Another way of extracting value from metadata is analysing it. Regulatory authorities are known for collecting large quantities of metadata and metadata can sometimes unique insights about the efficiency of processes in an organisation. For instance, dates and time metadata for regulatory actions can potentially be used to forecast activities.

As data grows and linkage becomes increasingly sought, metadata grows in importance inside an organisation, such that repositories of metadata become required to ensure that relationships between sets of data are adequately tracked and even historical relationships between variables are tracked (e.g. *patient_name* changed to *patient_first_name* and *patient_last_name*).

3.6. Opportunities in regulatory activities

As data size and variety increase, data manipulation at scale will become an important skillset in regulatory authorities.

Considering the time investment made in data manipulation, there is an opportunity to increase efficiency as well as capability by sharing data manipulation approaches within the network.

3.7. Challenges in regulatory activities

There are two major challenges in the implementation of transparent, robust and replicable data manipulation methods.

The first relates to skillsets. It is unclear how many scientific staffs in the network have the required skillsets in data manipulation, but this number is likely to be lower than what will be required going further. Training activities will likely need to be planned to ensure critical capacity in the network to perform big data analysis.

The second challenge is technology. Technology to use big data comes at a cost, both in terms of hardware and software. Implementing open source software will allow the widest possible application of methods across the network and ensure a commitment to open science and transparency with the additional benefit of avoiding some Member States investing in prohibitively expensive software. Hardware, on the other hand will be required. A rule-of-thumb with regards to big data is that big data is data that is bigger in size than the random-access memory of a computer and data sizes in routine analytics are now often at above personal computer capacity.

3.8. Recommendations

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
1	Data type	Actively checking and assigning data types	Report data type formats (i.e. analyst should attribute data type formats to avoid automated interpretations from software). Actively checking and assigning data types both after data collection and after importing data source.	Increased reporting and checking of data types
2	Reshaping data	The reshaping steps should be auditable, particularly where it concerns big data	Follow literate programming rules (i.e. analyst should describe what the code does) and retain an audit-trail. Use a consistent style guide. Protocol and track actions taken and rationale regarding data transformation, missing data and erroneous data. The analytical database table(s) created from reshaping should be saved as a separate data set.	Store and sharing programming codes with descriptions on the steps taken
3	Software	Consider the use of widely available open source software	Avoid or reduce the use of spreadsheet type software for data manipulation. Consider the use of widely available open source software. Widen possible application of methods across the network. Ensure a commitment to open science and transparency. Reduce the use of multiple software in the same analytical pipeline to reduced undetected errors.	Increased used of open source software

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation
4	Metadata	Document metadata thoroughly	Consider creating a repository of metadata and explore it to better unlock value from the data itself. Repositories of metadata become required to ensure that relationships between sets of data are adequately tracked and even historical relationships between variables are tracked.	Increased documentation of metadata
5	Governance	Sharing data manipulation approaches	Sharing data manipulation approaches within the EU Regulatory Network to increase efficiency and capabilities and enhance transparency.	Increased collaboration and sharing
6	Training	Plan and deliver data manipulation training sessions	Plan and deliver data manipulation training sessions for the EU Regulatory Network to foster know-how and capabilities to support big data analysis.	Increased capability in data manipulation of scientific staff

3.9. Resources

The following is a non-exhaustive list of resources for practical implementation of the above that the reader can use. A mature data analyst may choose to use online question and answer sites, software repositories or knowledge sharing sites, such as stackoverflow, github, dev.to, towardsdatascience.com etc.

Type of resource	Python	R	SAS
(online) Data manipulation in R https://www.datanovia.com/en/courses/data-manipulation-in-r/ https://itsalocke.com/files/DataManipulationinR.pdf		•	
(online) Data manipulation in SAS https://newonlinecourses.science.psu.edu/stat481/node/8/			•
https://www.oreilly.com/library/view/data-manipulation-with/9781785288814/		•	
(Books) Data manipulation in Python https://www.oreilly.com/library/view/python-data-analytics/9781484209585/ https://www.oreilly.com/library/view/python-for-data/9781491957653/	•		
(online book) pandas: powerful Python data analysis toolkit https://pandas.pydata.org/pandas-docs/stable/pandas.pdf	•		
(online) Coding systems for categorical variables in regression analysis https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/		•	
(online) Multiple resources on data manipulation https://stats.idre.ucla.edu/other/mult-pkg/seminars/		•	•
(online) Dealing with missing values in Python https://towardsdatascience.com/the-tale-of-missing-values-in-python-c96beb0e8a9d https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b	•		
(online) Dealing with missing values in R http://uc-r.github.io/missing_values https://www.statmethods.net/input/missingdata.html https://cran.r-project.org/web/packages/finalfit/vignettes/missing.html		•	
(online) R tidyverse style guide https://style.tidyverse.org/		•	
(online) Google's R style guide		•	

Type of resource	Python	R	SAS
https://google.github.io/styleguide/Rguide.xml			
(online) Python style guide https://www.python.org/dev/peps/pep-0008/	•		

4. The impact of artificial intelligence on analytics in the regulatory setting

4.1. Why

It is tempting to assume that the ever-increasing availability of electronic health data will transform the science of medicine and the way healthcare is provided. It is, however, essential to remember that *data by itself does not provide value*[28]: for data to be useful, it needs to be *analysed* (see Figure 15), and the insight derived interpreted and acted upon.

Figure 15: The intended role of analytics

Analytics is an important tool for gaining insights and providing tailored responses[29]; it can be defined as the research, discovery, and interpretation of patterns within data to help answering questions in three areas[30] (see Figure 16):

Description: providing a quantitative summary of certain features of the data. Descriptive analyses are used to answer questions such as the proportion of patients with a certain condition in a health care database (e.g. proportion of patients with diabetes, or how a drug is utilised).

Prediction: mapping some features (called input) to other features (called output) in order to estimate future or unobserved data. Examples of analyses in this category range from quantifying simple associations to using sophisticated pattern recognition methods on hundreds of variables to identify, for instance, which patients is likely to have a particular condition or experience a specific event (e.g. predict which patients are more likely to die within a week).

Causal inference: drawing conclusions about a causal connection between an occurrence and an effect. Causal analysis goes one step further than investigating associations among variables; its aim is to answer questions about the (counterfactual) impact of a change in policy or treatment[31]. Examples in this category include analyses to study the efficacy and safety of drugs (e.g. estimation of the effect of receiving a vaccine on a particular disease incidence).

---

[28] Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-9

[29] A Brief History of Analytics By Keith D. Foote on September 25, 2018

[30] Hernán, M. A., Hsu, J. and Healy, B. (2019). Data science is science's second chance to get causal inference right. A classification of data science tasks. Chance 32 42–49. Hernan's view is, of course, an aspiration and, in divorcing the field of Data Science/Analytics from the body of tools that have been developed to perform analyses it avoids the question of whether these tools can in fact fulfil the required tasks. It is thus necessary to evaluate the tools we have with respect to specific tasks of interest

[31] Pearl J. (2010) An Introduction to Causal Inference. The International Journal of Biostatistics. 2010:6(2)7 Published online Feb 26 2010

Figure 16: The different types of analytics

Insights in any of those areas can be used for two main purposes:

Support and inform decision-making by providing empirical evidence that might be neither obvious nor intuitive, *reducing the uncertainty* around the decisions and helping to see the consequences.

Enhance efficiency, through better scalability of a wide range of tasks amenable to some automation with benefits in greater consistency and reducing errors and costs.

The role of analytics is becoming of increasing interest as a result of the opportunities and challenges created by the vast volume of healthcare data that rapid developments in technology have helped capture. Part of the reason for the increasing interest in analytics is also the expanding prominence of Artificial Intelligence (AI), the focus of this chapter, that following the trichotomy classification depicted in Figure 16 finds most of its applications on the *predictive* tasks[32].

4.2. Objectives

The objectives of this chapter are twofold.

Firstly, a brief clarification of some of the terms associated with AI will be provided. The growing recognition of analytics and AI has moved the field from one of primarily academic interest to a mainstream discipline, with the potential to reach out to a vast audience but it has also resulted in misconceptions and misinterpretations.

Clarifying some of the terminology, creating the basis for a consistent approach and common language to be used in the regulatory environment, may help both the layperson and the analyst and will have the additional benefit of demystifying some of these concepts.

The second aim is to present the opportunities and corresponding challenges, with reference to the regulatory environment, in applying the more modern and promising approaches of advanced analytics. It is beyond the scope of this document to provide specific guidelines and recommendations since these depend on the specific applications; the aim will be to *establish clear principles* that will enable the development of specific guidelines: principles lead, more detailed guidelines follow.

The topic is presented focusing on the main concepts, assuming the reader is a scientist or assessor looking for guidance, or a trained analyst seeking clarifications and a structured presentation of the main methodological elements to consider when using AI algorithms on big data. Technical details have not been reported in the main text, but the reader interested can find some of them in the Annex B or in the references provided.

[32] For discussions on the use of AI for causal inference, Athey, Susan. 2015. "Machine Learning and Causal Inference for Policy Evaluation." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 5–6. ACM

4.3. Introduction

4.3.1. Defining the main concepts

Even if terms like advanced analytics, data science, AI, machine learning, deep learning and cognitive computing are becoming more familiar, in addition to terms that are still part of the analytical jargon such as data mining, knowledge discovery, pattern recognition and others, there is still confusion on what the terms really mean and how they relate to one another[33]. This is hindered by the fact that no precise definition has been agreed[34]: as a result, these terms are often used interchangeably.

Discussing in detail all the different terms and their role in modern analytics is beyond the scope of this report; the focus will be on providing clarity on the definition of AI and its main components in order to set the scene to understand opportunities and challenges.

The following will help to define AI for the purpose of this report:

There is not yet an agreed standard definition of AI: ISO/IEC JTC 1/SC 42, the technical sub-committee responsible of standardization in the area of AI, is currently working on the standard ISO/IEC 22989 'Artificial intelligence -- Concepts and terminology'.

> Merriam-Webster defines AI as 'A branch of computer science dealing with the simulation of intelligent behaviour in computers. The capability of a machine to imitate intelligent human behaviour'; the English Oxford Living Dictionary provides the following definition: 'The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages'.
>
> These definitions, and others available both in AI courses and from expert groups[35], make reference to human intelligence[36], a concept too wide to practically define AI (e.g. is there an actual algorithm, statistical model, or computer that performs tasks not requiring some degree of human intelligence?) and to help understanding its applications, potentiality and limitations.
>
> For the purpose of this report, the focus will be on what is at the core of AI: the *models* that consist of the algorithms to identify patterns in vast amounts of data and offer possibly actionable insights.

All currently deployed AI systems are examples of Artificial Narrow Intelligence[37]. AI can be classified in different ways; according to the extent of human tasks performed the following two categories are defined:

Artificial Narrow Intelligence (ANI, also called applied or special purpose): it consists only of some features of human intelligence where the criteria of success can be specified as a set of conditions and it can be very effective to solve a dedicated problem.

---

[33] For an overview of the definitions of some of these terms A Brief History of Analytics By Keith D. Foote on September 25, 2018 in https://www.dataversity.net/brief-history-analytics/

[34] Combi C - Editorial from the new Editor-in-Chief: Artificial Intelligence in Medicine and the forthcoming challenges. Artif Intell Med. 2017 Feb;76:37-39

[35] See for instance the High-Level Expert Group on Artificial Intelligence - A definition of AI: Main capabilities and scientific disciplines. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf

[36] A definition of intelligence useful to better define AI can be found in The Measure of Intelligence François Chollet arXiv:1911.01547

[37] See for instance the High-Level Expert Group on Artificial Intelligence - A definition of AI: Main capabilities and scientific disciplines https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december.pdf

Artificial General Intelligence (AGI): it supports a broad range of solutions and tends to include a wider range of characteristic of human intelligence.

AI is a collection of methods and technologies: AI is not a single entity, defined by a single method. Rather, it is a collection of data science technologies in varying degrees of maturity; some have been around for decades (or more), others are relatively new. A short description of the methods with most potential and interest follows.

4.3.2. Two approaches to AI

The term AI was coined by John McCarthy in 1956 and includes two main areas of research. One is based on rules, logic, and symbols; it is explainable; however, it can be used only when all possible scenarios for the problem at hand can be foreseen.

An example in this area is a rule-based system[38], such as the one used for medical diagnosis where, for example, if a patient has a sore throat, runny nose, and mild fever, the patient probably has a cold. Another example is in the chess game. The first to face chess was Alan Turing in 1948, who created an algorithm to play chess based on logic. Many years later Deep Blue, based on heuristic search, was the first computer to win a human world champion facing Kasparov.

One of the most difficult challenges of this approach is to code or program in painstaking details the steps: how and in which order to do it. There is, however, another more promising approach: writing a very general program for how to learn from the data. Then a machine can be taught what needs to be done through providing it with lots of data and, often, telling it the right and wrong answers along the way. This is what is called *machine learning*[39], approach that became popular in the late 80s and 90s. Examples of applications in this area are algorithms to recognise human faces (Facebook), self-driving cars (Google Autonomous Cars) or the most recent developments in playing the game of chess[40].

This approach differs from the more traditional rule based, logic or symbolic techniques where the knowledge required to solve the problems have to be specified a priori in terms of statements and relations using some formal logic; instead learning programs will discover relationships among observed facts autonomously which enables the use of algorithms *in situations where it is not possible to code explicitly a solution or to model rules able to match every possible scenario* (see Table 4). These two approaches can also be combined in order to maximize the advantages of both and to mitigate their drawbacks.

Table 4 Difference between traditional programming and machine learning

Traditional programming	Machine learning
Pre-programmed by humans: producing the same results every time	*Adaptive*: changing its behaviour based on data and performance

[38] AI research is (unsurprisingly) divided on the issue on whether rule-based system can be considered AI. For some, the learning component is necessary, for others, the ability to simulate some degree of intelligence even without the ability to learn is enough. It is a definitional problem that does not impact the discussion of this report (focused on machine learning algorithms); in this report rule-based systems are regarded as the simplest form of AI.
[39] Artificial Intelligence: implications for business strategy. MIT Management Executive Education online course 2018.
[40] Deep Chess offered an end-to-end neural network implementation, and AlphaZero Chess has been implemented using Deepmind's reinforcement learning technology.

Traditional programming	Machine learning
Single purpose: for one/limited purpose	*Multi-purpose*: potential for more general purposes
Code driven: the solution is given by programming a sequence of deterministic instructions	*Data driven:* the solution is given by a model driven logic learned by induction from data

*source: Machine Learning Security, https://www.slideshare.net/eburon/machine-learning-security-ibm-seoul-compressed-version

4.3.3. Machine learning

Today, most of what is commonly called AI, has machine learning algorithms at its root (see Figure 17).

ISO defines machine learning as a "process using algorithms rather than procedural coding (explicit programming instruction) that enables learning from existing data in order to predict future outcomes"[41]. The algorithms can also model and improve its performance in response to new data and experiences to improve efficacy over time[42] (see Annex B for more details).

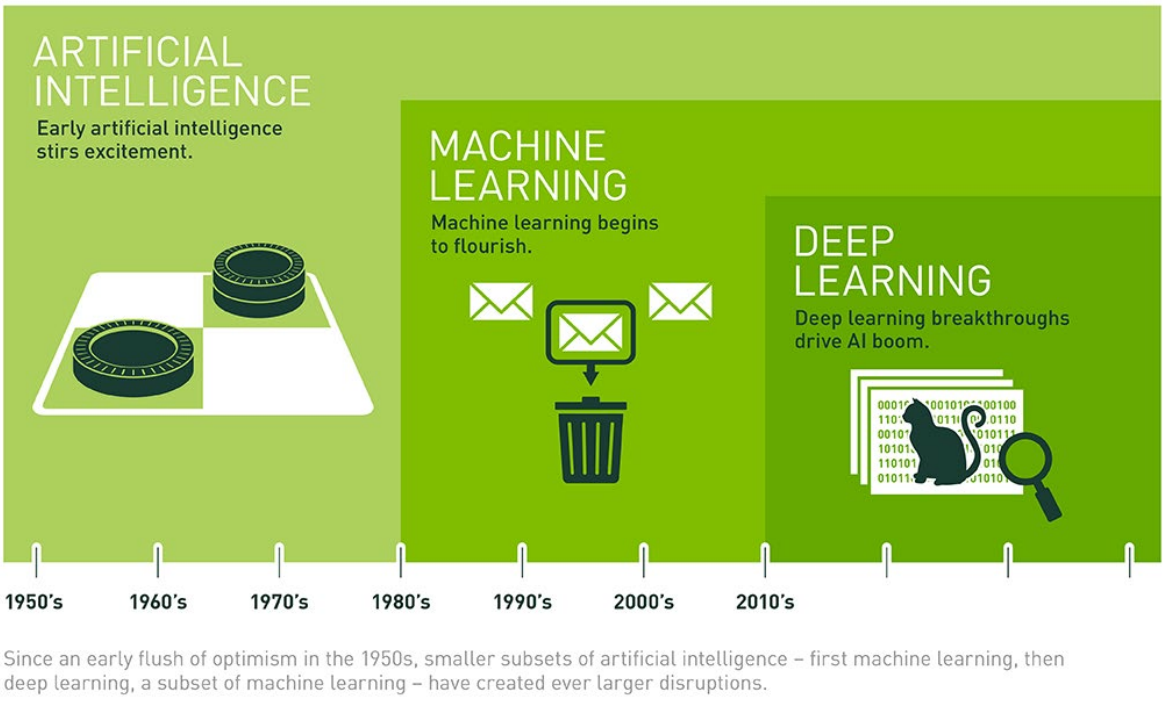


Figure 17: Relationship between artificial intelligence, machine learning and deep learning

[41] ISO/IEC JTC 1/SC 40
[42] Chui M, McCarthy B. An executive's guide to AI. McKinsey & Company 1996-2019.

Source: NVIDIA, https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Machine learning collects a plethora of techniques and models, some of them may sound familiar to who has a background in data analysis and statistics; linear or logistic regression, k-means clustering, decision trees, random forests, Bayesian networks, just to cite a few.

Machine learning can be used to build models concerning data with different nature and scope. One particular model which exploits the richness of longitudinal healthcare data and hence is particularly relevant to medicines regulation is the *time series model*: a sequence of observations on variables that are measured over successive points in time. As such, an increasing number of scientific publications are exploring the use of time-series in real-world data in the context of drug regulation (see Annex B for more details).

4.3.4. Deep learning

Deep learning is one family of models belonging to machine learning. In the modern era of availability of big data and computational powers, they are currently considered the most promising; its use in self-driving cars or to automate complex tasks such as recognizing voice commands, or supporting complex simulations in drug discovery, made these techniques very popular.

Deep learning can be regarded as an advance in the field of artificial neural networks (ANNs), models loosely inspired to biological neural networks, formed by many interconnected functional units called neuros.

ISO defines neural networks as the "computational model utilising distributed, parallel local processing, and consisting of a network of simple processing elements called artificial neurons, which can exhibit complex global behaviour"[43].

Deep learning takes the idea of ANN further, by building complex neural architectures, where each block (generally a layer of neuros) is aimed at focusing on a specific part of the model. Each block receives the output of the preceding blocks and feed following blocks; this process helps to progressively detect features and complex patterns in the input data. For instance, in the case of deep learning applied to image analysis, a first layer can be tuned to detect lines or simple shapes, the following one starts combining them to form more complex shapes and so on. It is called deep as the number of layers can go into the hundreds[44] (see Annex B for more details).

The strength of these kinds of models resides in the following:

They tackle complex problems, being able to learn complicated patterns from large high-dimensional and heterogeneous data[45] and from previously intractable data types (such as images, speech and video) which has appeal for healthcare data applications.

They tend to be data-driven and autodidactic; usually raw data can be fed into the model without much data transformation[46].

[43] ISO/IEC JTC 1/SC 40
[44] Any neural network of more than just two layers can be called deep learning (Patterson & Gibson 2017: Deep Learning)
[45] Geron A. Hands-on-Machine Learning with Scikit-Learn & Tensorflow. Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media 2017 Mar
[46] Beam A.L, Kohane I.S (2018) Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317-1318 published online April 3 2018: https://jamanetwork.com/journals/jama/fullarticle/2675024

4.3.5. Natural language processing

Data generated from conversations, articles, medical notes or even tweets are examples of unstructured data. These data do not fit neatly into the traditional row and column structure of relational databases, represent the vast majority of data available in the actual world and are usually not curated and hard to manipulate[47]. Natural language processing is the branch of AI that enables machines to read and process this kind of data.

ISO defines natural language processing as the "analysis by computers of text in natural language to gain meaningful information from human language" [48].

Sometimes natural language processing tasks are relatively simple, like for instance tokenising a text to be able to compare it to another. For more complex tasks, currently natural language processing models are battling to detect nuances in language meaning due to the inherently un-curated nature of unstructured data (e.g. spelling errors, dialectal differences, and context dependent). Deep learning seems currently the most promising approach in this field[49].

There are three ways in which natural language processing can derive value from unstructured data, either in text or in speech: i) analysing them, ii) generating them and iii) putting the two together to have a conversation.

4.3.6. Why AI is becoming popular

While many of the ideas and technologies related to AI have been around for decades, they have gained in popularity recently due to the convergence and interrelationship between three key components:

Abundance of collectable *data*: real time flow of all types of structured and unstructured data from social media, communications, digital imaging, sensors and devices are now available, in addition to a wide range of data from many other research initiatives (e.g. omics projects, digitization of health-related records).

Advances in *computational power* (e.g. highly scalable cloud-based architecture and use of massively parallel processors, e.g. Graphics Processing Unit (GPU) and Tensor Processing Units (TPU)) and *storage* along with their availability and affordability.

Advances and availability of *algorithms and methods* and in particular the availability of deep learning. New and refined methods are diffusing rapidly through a combination of open-source and proprietary programs sponsored both by technology giants and many academic and non-profit teams[50]. The development of freely shared algorithms and software has driven, in turn, the availability of accessible and extensive training and community-based support (e.g. Tensorflow, Pytorch, Keras).

Most of the advances in the field derive from an increase in the availability of data, computational power and algorithms and not by the development of new algorithms despite the hype around them.

---

[47] Lopez YD. An Introduction to Natural Language Processing (NLP) OpenDataScience.com 2019 Jan.
[48] ISO/IEC JTC 1/SC 42/WG 1
[49] Artificial Intelligence: implications for business strategy. MIT Management Executive Education online course 2018
[50] Factors Driving Adoption of AI and Deep Learning - Seven factors appear to be driving the rapid uptake of AI and deep learning

4.3.7. Aim of the AI algorithms

It has been mentioned as AI applications are mainly for predictive tasks. Within this category, a further distinction can be made based on the focus of the model: solely on maximising the *prediction capabilities,* creating 'black-box' models, or also on *inference/interpretability* to understand how the model generates its estimates[51].

Prediction: in many situations, a set of variables denoted as inputs (or features or predictors) are readily available and are thought to have some influence on one or more outputs (or events) that cannot be easily obtained or measured. Predictions are estimates of future or unobserved states based on observed data[52]; however, they are not the same as causal relationships.

Inference[53]/interpretability: at other times the main interest is in understanding how the output is associated with the input/features. This setting is useful when trying to answer the following questions:

Which input variable is associated with the output? It is often the case that only a small fraction of the available input variables is substantially associated with the output of interest?

What is the relationship between the output and each input variable? Some features may have a positive relationship with the output; others a negative. Depending on the complexity of the model used, the relationship may also change according to the values of the other features?

How can the relationship be summarised, is a linear model adequate or a more complicated relationship is needed?

How observations or variables are related to each other: the aim in this case is to discover the inherent groupings in the data, for instance to reduce the number of variables or observations whilst ensuring that important information is still conveyed?

This difference between prediction and inference/interpretability can be highlighted with a simple example: is the aim to predict the subgroup of patients for whom the treatment is beneficial or to understand what makes this subgroup different from others to inform the development of future medicines?

As discussed subsequently, depending on whether the ultimate goal is prediction, inference, or a combination of the two, different methods may be more appropriate. For example, linear models *facilitate inference* in terms of human interpretability, but may not yield as accurate predictions as some other approaches. In contrast, some sophisticated non-linear approaches can potentially be more accurate, but this comes at the *expense of interpretability*.

---

[51] Hastie T, Tibshirani R, Friedman J. Springer 2016. The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition
Machine learning emphasises more prediction over inference which is what statistics is concerned about. For the readers more interested in the difference between machine learning and traditional statistics: Hassibi K. Machine learning vs. traditional statistics: different philosophies, different approaches. DataScience Central.com 2016 Oct.

[52] Prediction is a general problem of what to do when we do not have enough information. It isn't necessarily about the future

[53] It is worth noting that inference is used in a rather wider sense in general statistics. The way it is used in this report serves to highlight the difference between where the focus of the prediction model is: only prediction or also understanding what variables are associated with the output to predict. Moreover, it should not be confused with the concept of causal inference described in the introduction

4.3.8. Which AI algorithm to use

There are many different types of machine learning algorithms[54] rather than just a single best method. The reason lies in the fact that there is no a single method that dominates over all possible *data sets* and all possible *applications*. For each particular method there are situations for which it is particularly well suited and others where it does not perform as well. These situations are seldom known in advance and selecting the best approach can be one of the most challenging parts of performing advanced analytics in practice[55].

Explaining in detail each algorithm, its strengths, weaknesses and applications is out of scope of this report. However, in this and the next section several principles will be defined to inform the choice and acceptability of the best algorithm for a particular situation.

Most of the hype around the more sophisticated machine learning and deep learning algorithms is about how they improved the accuracy of predictions; but this increase is not obtainable in every situation. Deep learning architectures are specifically designed to solve one problem, such as object recognition in images or translation between languages, their applicability to different problems may result in lower or even poor performances; moreover, deep learning generally requires a large amount of data to train the model, thus the need of using big data.

However, it is reasonable to believe that non-linear relationships are likely to be better modelled when an appropriate non-linear model can be selected, and it is known that some machine learning methods can adapt to a wide variety of such relationships. These methods are also well suited to take advantage of the high dimensionality of the data (being able to automatically process and model a high number of variables).

The challenge in statistical interpretation of such models is that more degrees of freedom are used in fitting a more complex model and so the extent to which the fit of the model will generalise to data beyond that used to fit the model cannot usually be calculated from theoretical considerations.

4.3.9. Summary of the main points

- Most of what is commonly called AI, has machine learning and deep learning algorithms at its root.
- Though the terms machine learning and deep learning gained recent popularity, many of the concepts and models that underlie it were developed decades ago[56].

While the level of sophistication of the methods has increased, the objectives remain the same; to find patterns in vast amount of data. A deep learning of today has at its core finding correlations in the data as a linear regression of the past[57].

- However, those patterns may be more opaque or less discernible to find, in data that previously was not easily analysable (e.g. unstructured or images), and therefore *now there is the capability of representing more complex phenomena.*

---

[54] An introduction to the array of different algorithms and the potential and pitfalls of supervised learning can be found in: Mullainathan S, Spiess J. Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives. Volume 31, Number 2. Spring 2017. Pages 87–106

[55] Gareth J, Daniela W, Hastie T, Tibshirani R. Springer 2017. An Introduction to Statistical Learning: with Applications in R

[56] The earliest form of what is now known as linear regression was developed at the beginning of the nineteenth century and even the newer and more promising methods as deep learning are based on techniques that have been largely developed in the middle of the twenty century

[57] Kalev L. Does AI truly learn and why we need to stop overhyping deep learning. Forbes.com. 2018 Dec.

Moreover, the quantity, complexity and high dimensionality of data available can be exploited by models with a high degree of flexibility in order to enable better predictive accuracy.

It is wrong to assume that all scientific fields of study, all research questions and all predictions are better explained or should always be modelled using the most sophisticated machine learning techniques. While the variety and speed of growth of machine learning methods provides an irresistible attraction towards ever-complex models, the principle of Occam's razor[58] should be firmly in the mind of the data analyst. To paraphrase Einstein "make your model as simple as possible, but not simpler".

4.4. Opportunities in regulatory activities

AI may offer opportunities to improve regulation, either by being *more efficient* and freeing resources for core regulatory activities or by *enhancing support for regulatory science and decision making*.

Taking advantages of these opportunities requires a change in perspective in how tasks are viewed. It has been argued that if a task can be framed as a prediction problem with well-defined input and output, even if the process in between is not fully described, the use of machine learning can be applied with a significant/substantial impact[59]. And when machine learning models are framed as cheap prediction, their potential becomes clearer: *prediction is at the heart of making decisions under uncertainty*, both operational and strategic, and *prediction tools increase productivity*.

Following this structure, the potential applications of machine learning will be described in two categories:

Process automation: applications where the main aims are efficiency, time and/or resource saving, increased robustness and consistency for processes, and reduced human errors. This will usually apply, but not only, to repetitive administrative/technical tasks.

Enhance regulatory science to support decision-making: applications where the main aim is to provide more insights reducing the uncertainty around options.

This classification contains some overlap as some applications will have an impact on both categories, but it can be a useful starting point for the description that follows.

An advantage that will permeate both areas is that the use of algorithms will provide more consistency (mitigate inter observer variability) compared to when humans are involved. The potential penalty involved is that apparent lack of consistency in human decisions may reflect subtle differences in the context that has not been captured by an automated system or by the data used to train it.

Finally, some of the examples and case studies reported do not directly link to the regulatory environment (i.e. decision-making in medicines development and approval), yet they are described as they might have an indirect beneficial effect, they might be relevant as early awareness for future opportunities, and they all contribute to promote public health.

[58] Occam's razor is a principle from philosophy; suppose two explanations for an occurrence exist, the simpler one is usually better. Or, the more assumptions are needed, the more unlikely an explanation is

[59] Hunt S. From Maps to Apps: the Power of Machine Learning and Artificial Intelligence for Regulators. Financial Conduct Authority. 2017 Nov.

4.4.1. Efficiency and automation

AI can play a meaningful role in making administrative processes faster and more efficient; any manual task that is repetitive and have relatively clearly defined outcomes should be a candidate to be streamlined.

However, medicines, and healthcare in particular, are very error sensitive fields. Every potential algorithm needs not only to be tested extensively before being in use but also monitored whilst in use; it is also likely that a human component will still be needed. In some applications a *semi-automatic* approach may be preferred where humans perform conflict resolution and carry out compliance checks, or where AI makes suggestions only and final decisions are made by humans; this may also be more efficient overall compared with a fully automatic application. This involvement would not only be useful in ensuring a high accuracy of the algorithm over time but would also provide valuable feedback that can be used by the algorithm itself for improvement.

The point made implicitly is that applications using AI will not be 100% accurate; there will be an error rate. The consequence is that for every application an acceptable error rate should be defined in advance in accordance with the sensitivity of the task (e.g. it is more important to avoid false positive or false negative[60]) and the main aim of the initiative (e.g. if the main aim is saving resources a slightly higher error rate could be acceptable). In case the AI application cannot guarantee such an error rate, the semi-automatic approach with a human component may be the preferred approach. Usually error rates are not the same across all scenarios and steps in the process; as such the steps where the error rate is higher could then be done in conjunction with or completely delegating to humans. For instance, while testing AI to automate quality checks, if it is seen that a particular type of check has a higher error rate, this check could be delegated to a human while all the rest could be done by the AI application.

The most promising applications in this area are:

Chatbots for provision of safety information or to complement / augment service desk: this could have different flavour, from a simpler automatic call deflection (assigning the request per topic and/or complexity to the right team) to use historical data to create a knowledge-based repository and answer the requests received.

Quality assurance frameworks: assuring and checking the quality of the data collected or submitted is an important work to ensure the validity of any evidence generated with that data, but at the same time it is a tedious work that often absorbs a considerable amount of resources.

> Using AI to automate at least a part of the quality checking would increase the speed and consistency of this activity; whilst at the same time would contribute to saving resources. The possibility to use the vast amount of historical data with the corresponding checks makes it a good candidate for machine learning algorithms.
>
> An example would be the validation of the authorised product information for eXtended EudraVigilance Medicinal Product Database (XEVMPD) database EMA receives from Marketing Authorisation Holders (MAH) in the form of structured Extensible Markup Language (XML)

[60] Borrowed from medical testing, in binary classification, false positive / false negative is an error due to the misclassification of a data point, in particular when it is erroneously attributed (false positive) and when it is erroneously not attributed (false negative) to a class. In statistical hypothesis testing the analogous concepts are known as type I and type II errors

messages. This structured information is populated by MAHs based on the Summary of Product Characteristic (SmPC) unstructured document.

After receiving XEVMPD data, EMA outsources the validation of messages versus the original SmPC to external providers. A machine learning algorithm could be trained to validate XEVMPD XML messages against unstructured SmPC and EudraVigilance controlled vocabularies and could propose conflict resolutions to EMA experts for confirmation.

Existing SmPCs, XEVMPD XML messages, and the historical decisions made by external contractors and EMA experts on conflict resolutions provides a unique and constantly growing training data set.

A point to consider is that quality assurance of data to detect malpractice is a particularly challenging area since it would be expected that methods used to produce the data would adapt over time to evade the capabilities of the AI process. Thus, it may be preferable to avoid reliance solely on automated data quality checks when malpractice is considered a possibility.

Automation of processes: this category includes a very heterogeneous set of examples such as the generation of Individual Case Safety Report (ICSR) or XEVMPD messages, narratives, redaction of information to be made public, etc. All these examples share the benefit of automation highlighted previously and the possibility to exploit the vast amount of historical data.

Literature monitoring: possibility to automatically search and screen for literature articles based on specific research questions or to support routine monitoring of information given the arduous task to keep updated in the proper area of research.

An example would be the monitoring of literature for articles concerning Adverse Drug Reactions (ADR) and the corresponding creation of ICSR messages. Searching for articles containing substances of interest and screening to check that they also contain sufficient information concerning ADRs, whilst at the same time meeting the inclusion criteria to be reported, could benefit from natural language processing and machine learning algorithms

- *Variation validation*: create an AI algorithm that evaluates the variation type (Type 1A, 1B, 2, etc.) and analyse whether a submitted application is correct in type and format flagging where possible errors or inconsistencies are present.
- *Personal data redaction*: an AI algorithm that recognises and redacts personal data in documents that needs to be published or delivered.

One interesting example not directly belonging to the regulatory field follows:

Streamline routine work by health care professionals: different sources offer varying estimates of the amount of time spent by health care professionals on tasks amenable to some automation and research studies also suggest specific possibilities for reduction in errors and improved workflow in clinical settings with appropriate deployment of AI[61].

4.4.2. Support regulatory science and decision making

Even choosing where to focus and setting out priorities can be considered a prediction problem where AI algorithms can support prioritisation decisions by efficiently taking into account all the data available.

[61] Naylor D. On the Prospects for a (Deep) Learning Health Care System. JAMA Network. 2018 Sep. https://jamanetwork.com/journals/jama/fullarticle/2701667

AI has a wide range of potential applications in providing insights to support decision making in the hypothesis generation phase. This latter use varies in importance across different regulatory and drug development functions. For surveillance of drug safety or identifying promising areas for extensions of indications, the abilities of AI to find relationships in data appear ideally suited. However, the capabilities of AI applied to observational data to provide measures of certainty for use in formal decision-making processes comparable to those obtained from designed experiments requires further investigation.

Drug discovery process: examples include sophisticated natural language processing searches of biomedical literature, data mining of millions of molecular structures, designing and making new molecules, predicting off target effects and toxicity, predicting the right dose for experimental drugs, and developing cellular assays at a massive scale. There is new hope that preclinical animal testing can be reduced via machine-learning prediction of toxicity[62] and that more in silico exploration will precede in vitro examination and in vivo experimentation[61].

Selection/manufacturing of individualised treatments: AI algorithms are used for mutation detection by exome sequencing, selection of vaccine targets by solely prioritisation of mutated epitopes predicted to be abundantly expressed and good MHC class I or class II binders[63] .

Matching and recruitment of patients to clinical trials based on specified criteria: analysing structured data, such as ICD-10 or SNOMED codes, and unstructured clinical data, such as doctors' notes and medical data in free text format, in order to convert fragmented medical documents into the information needed to match complex clinical trial criteria. This could target specific patient groups and diseases that, in turn, would lead to savings in the form of fewer test subjects, test subjects more likely to benefit from the treatment, shorter experimental periods and fewer trials to be changed or abandoned.

Diagnosis: a quicker and more accurate diagnosis avoids the risk of errors and allows a faster time to treatment.

> Using machine learning algorithms to integrate information on signs and symptoms presented by individual patients (including diagnostic tests and images) is showing promising results in identifying diseases. In one study which aimed to identify possible patients with a rare disease, machine learning algorithms performed 2.5 times better than a standard epidemiological approach and 5 times better than clinical expert review[64]. It is expected that the performance of deep learning and other machine learning methods will improve even further with exposure to more data and the availability of more linked data sets. An example in this area is Apache cTAKES™, a service that uses natural language processing to extract disease conditions, medications, and treatment outcomes from patient notes, clinical trial reports, and other electronic health records[65].
>
> This capability has numerous potential applications, for example to:

Better understand determinants and clinical progression of diseases (and other events).

Identify the frequency of undiagnosed patients and estimate unmet clinical needs.

---

[62] Topol E. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicines 25, 44-56. 2019 Jan.

[63] See the 'Bioanalytical Omics' subgroup report for more details

[64] Presentation by Rigg J. Using machine learning and real-world data to tackle complex healthcare challenges. IQVIA April 2018

[65] Lopez YD. An Introduction to Natural Language Processing (NLP) OpenDataScience.com 2019 Jan.

Create potential cohorts of patients eligible to enter a clinical trial or observational study.

Predictions of future outcomes and improve efficiency of the treatments: the use of AI algorithms to:

Estimate more accurately future drug effectiveness based on analysis of randomised controlled trials efficacy data and historical data on compliance, dosage, and concomitant use of other products.

Identify prognostic factors/biomarkers to improve targeting, for example to determine adherence or cases of abuse.

Improving therapeutic decisions by matching people to the best treatments based on their specific health, life-experience and genetic profile (i.e. personalised medicine).

Earlier and more accurate identification of adverse drug reactions: faster, more effective discrimination of signals from noise, allowing a better scalability of the current processes (see Spontaneous ADR group).

> Advanced analytics could take full advantage of the richness of information available in databases of spontaneous reporting. Currently, the more common algorithms used to prioritise the drug-event association that may represent a safety problem use only a fraction of the information available. In the future, use of more of the information already collected in spontaneous reports and linking information from other datasets like pharmacological information about the medicines[66] could deliver a faster and more accurate identification of safety signals.

Detected changes in trend: time series models can be used to detect an unexpected rise in the frequency of an adverse drug reaction that may indicate the presence of a quality defect, medication error or abuse/misuse; none of these is easily detectable with commonly used analyses. Similarly, the trend-change in a time series after the implementation of a risk minimisation measure may give an indication of the impact of the measure.

Extract structured information from various data sources, such as ICSRs, literature articles, product information, medical charts, free text fields of electronic health care records, images or social media.

> This application starts with using natural language processing to automate processes that are currently performed manually. The output can then be used to support decision making by creating metadata repository that can either be i) searched; ii) monitored via dashboards; or iii) serve as input for more advanced analytics.
>
> For instance, the creation of a structured repository of the information contained in the SmPC would facilitate the following use cases:

The creation of a database of adverse reactions reported in the SmPC that can then be used to inform the signal detection process.

The possibility to search for all products with a particular indication (e.g. all products authorised in the paediatric population) and monitor their evolution over time.

The identification and stratification of risk factors for the occurrence of specific ADRs.

The possibility to link with other databases (e.g. drug structure, etc.) and using advance analytics as, for instance, uncovering trends, estimating unmet medical needs, clustering products according to similar safety profile, indications, etc.

---

[66] Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-Driven Prediction of Drug Effects and Interactions. Sci Transl Med. 2012 Mar 14;4(125):125ra31

Other examples in this category relative to clinical trials are:

The creation of electronic case report forms extracting data from medical charts. A review of medical notes in the US found that medical charts contain 75% of unstructured data.

In phase IV clinical trials with randomised patients in treatment arms and data collected on clinical outcomes in GP-based electronic health records, the extraction of relevant information on outcomes from the free texts' fields.

Identify duplicates in databases: algorithms that take full advantage of the richness of the data available would be ideal candidates for identifying duplicates.

Other noteworthy examples not directly belonging to the regulatory field are:

- Natural language process can be also used to recognise and predict diseases based on the *patient's own speech* in a variety of health conditions from cardiovascular disease to depression, and improve in the treatment of Alzheimer's disease by monitoring cognitive impairment through speech[51].
- Clinical application of deep learning has been the most rapid in image-intensive fields such as radiology, radiotherapy, and image-guided surgery. In many cases, interpretation of images by deep learning systems has outperformed that of individual clinicians[61]. Currently images are mostly used for binary decisions like presence of a disease or progression or not, but it is likely they contain much more information than that.

4.5. Challenges in regulatory activities

With the current excitement around AI, it is easy to ignore the associated limitations and risks, in particular considering the exploratory approach inherent in AI. However, if these limitations are understood, they can be managed and mitigated, and the potential of AI can be better exploited.

As already mentioned, the main aim of analysing such a rich and heterogeneous set of data is to improve efficiency and to support the evidence generation on the effectiveness, harm and value of medicinal product to inform decision making. Focusing on the latter, Schneeweiss[67] defines the *key evidentiary needs* to enable successful regulatory decision-making in four broad requirements; these have been adapted considering the focus of AI on prediction and are useful, even if with different degrees, when utilising AI for automation:

Meaningful evidence that provides relevance and context sufficient to inform the interpretation and draw conclusions to support decision-making.

Valid evidence that meets scientific standards to allow predictive accuracy.

Expedited evidence that provides evidence synchronized with the decision-making process.

Transparent evidence that is audible, reproducible, robust, and ultimately trusted by decision makers.

A major challenge is to decide if these needs can be translated into *requirements for regulatory acceptability* that an AI algorithm should satisfy; when this is possible these requirements will in turn highlight *where attention should be paid to* by either MAHs or assessors while using or assessing analyses provided through AI algorithms.

[67] Schneeweiss S, Glynn R. Real-World Data Analytics Fit for Regulatory Decision-Making. American Journal of Law & Medicine, 44 (2017): 197-217

Meaningful evidence: the data element

In order to be meaningful, the evidence generated needs to contain *relevant* information of *adequate* quality to answer the desired question: this requirement is related to the data more than the analysis, but it offers a good opportunity to reaffirm that even in the era of AI the concept of '*garbage in – garbage out*' still applies[68].

Data are at the core of any algorithm since they are *used to train them*. The abundance of data and newer algorithms cannot automatically compensate for missing data and other data biases in the observational dataset; data pitfalls can still substantially affect the results and, as some authors reported, *data is still more important than algorithms*[69].

The core challenge is that most big data that has received popular attention is not the output of instruments designed to produce valid and reliable data amenable for scientific analysis[70].

To avoid these biases, the following areas should be critically assessed:

Quality of the data: the quality of the data used to train the model has to be examined in term of consistency (e.g. outliers/extreme or illogical values, etc.), completeness (proportion and cause of missing values, etc.) and especially accuracy (e.g. measurement error). For instance, outlier or illogical values can have a disproportionate weight on the model and heavily influence its results, errors and or biased human judgements present in the data used to train the algorithms will be reflected in its results.

Representativeness of the data where the model will be used: data not representing reality is a very common problem, either because of data collection methods or because of actions that truncate the data (e.g. privacy rules, limited inclusivity, etc.). One example is the corpus of genomic data, which so far has seriously underrepresented minorities: 'without inclusivity, there cannot be a democratisation of health' [71]. Another example is when algorithms are trained on historical data to inform future predictions or inference: for instance, if the staging of a cancer is not the same before and after the PET-CT scan era, the historical data do not represent the current patients.

Valid evidence: the methodological element

After considerations of the validity of the data, the next step is to be able to produce a valid analysis. In this regard, it should be considered that *the statistical lessons of the past should not be forgotten* while embracing new data and methods[69]. Sophisticated AI algorithms may be good at predicting outcomes, but predictors are different from causes. The usual common-sense caveats about *confusing correlation with causation* apply; they become even more important as millions of variables and observations are included in statistical models[72]. It should always be remembered that AI does not solve any of the fundamental problems of causal inference and bias in observational data sets; causation has not been "knocked off its pedestal" [69].

[68] In analysis and logic, "garbage in, garbage out" (GIGO) describes the concept that flawed, or nonsense input data produces nonsense output or "garbage"

[69] Harford T. Big data: are we making a big mistake? FinancialTimes.com 2019 Mar.

[70] Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science 2014 Mar. Vol. 343, Issue 6176, pp. 1203-1205

[71] The Democratization of Health Care. Stanford Medicine 2018 Health Trends Report. 2018 Dec.

[72] Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-9

Causality versus correlation: all statisticians' tools applied to observational data are for analysing correlations. They are entirely silent about causality so that the causal interpretation must come from somewhere outside the model.

> The goal of an analysis may be to *discover causal associations in order to make accurate predictions*, but the optimisation objective for most AI models is simply to *minimise errors*, and this might be achieved just discovering associations[73]. Some of these are confounders, some others are causal components. The Google's flu example helps highlighting this point: Google's engineers did not aim to figure out what caused what; they were merely looking for statistical patterns in the data. They gave more importance to correlation rather than causation, to prediction accuracy rather than to understand the relationship in the data; a common trend in big data analysis. However, if one has no idea what is behind a correlation, there is also no idea about what might cause that correlation to break down.
>
> Validating that an association is causally related, or at least the mechanisms by which predictions are made, is key in generating evidence to support regulatory decision making; while it is less important when using algorithms to automatise tasks.

Precision and accuracy: big data via the volume often tends to increase precision of results (moving from left to right in Figure 18) but does little to address the accuracy or validity (moving from top right to bottom right in Figure 18). This could lead to the dangerous situation where both very precise and consistently biased results (top right quadrant in Figure below) can be obtained. A highly precise biased result, especially if perceived as credible based on precision alone, is more dangerously translated into practice than an imprecise biased result[74].

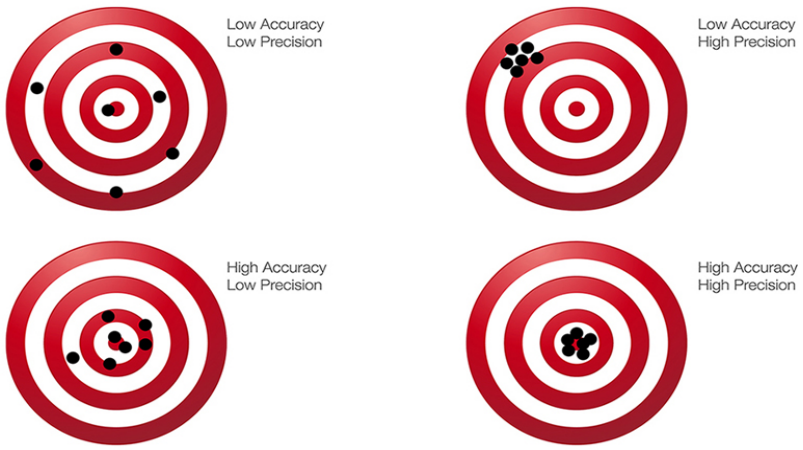


Figure 18: Precision and accuracy

Multiple comparisons: the large number of choices involved in any retrospective study is a major statistical issue. Multiplicity and the complexity of the data could be exploited to produce false results that would be difficult to detect test enough different hypotheses and it is plausible that some of the

[73] Hastie T, Tibshirani R, Friedman J. Springer 2016. The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition

[74] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. 2017 Apr 27;9:245-250

results are due to chance. This is more serious in large data sets, because there are vastly more possible comparisons available for testing, and when using more flexible algorithms, where the great number of choices regarding their parameters is in effect similar to testing large number of alternative hypotheses. For instance, deep learning methods have been accused of 'cherry-picking' since large amount of parameters / hypothesis are tested.

Generalisability of the model (or external validation): the main assumption when constructing machine learning models is that results will *generalise beyond the data used* to fit and train the model; then it would be justified to use the model on unmeasured observations.

Some of the machine learning models available are very flexible and can easily provide a nearly perfect result on the data they have been trained on; however, they will also provide less accurate predictions when used on new observations of real-world data. This phenomenon, which may seem counterintuitive at first glance, is called overfitting.

The reason behind the poor performance on new observations is that, when the parameters of a model have been tuned so that it perfectly fits the available data, the model captures both the intrinsic relationships between the variables, and the noise in the data. The consequence is that over-fitted models do not generalise when used with new data, and thus the results are not useful. Specifying and training models is much about balancing the *trade-off between the accuracy of the model and the generalisability* of its results. Techniques to avoid overfitting are called *regularisation techniques* and this currently an area of intensive research.

This generalisability concern is severe and must be addressed by testing models on truly independent validation data sets, from different populations or periods that played no part in model development. If external data are not available, a common approach is to divide the available data randomly into training and test sets, at an 80:20 split for instance. The model is then created using just the training data and then evaluated on the test data. It is thus essential that the test set resembles the data for which predictions are going to be made. Cross-validation approaches elaborate on this concept using a resampling procedure that allows using all the data in testing and training; it is considered 'superior' to the simple split because it produces more unbiased results (see Figure 19). Regulatory monitoring of these data splitting activities to ensure that the data are divided independently of the results of any analysis is very challenging.

Figure 19: K-fold cross validation approach

Source : https://sebastianraschka.com/faq/docs/evaluate-a-model.html

This principle is so important that in many data science competitions, validation data are released only after teams upload their final algorithms built on another, publicly available data set.

- *Performance (or accuracy)*: the model's overall accuracy or predictive power is usually the first performance metric by which a model is evaluated, and it is assessed by estimating the error on the test database. Good performance in this area has been one of the main drivers in the continuous fine tuning of machine learning models and in the interest they have generated.

  When assessing the performance of a model, there are typically many possible outcomes and it is unlikely that an algorithm will forecast every one of them precisely[75]: it is important to identify which ones are most likely to lead to the best decision for the specific situation. The performance measures to consider depend on the type of model used (e.g. supervised versus unsupervised learning, discrete versus continuous outcome variable) and on the context where the model is applied. In some cases, as for instance in identifying potential safety signals, having a higher sensitivity could be more important, however this statistic alone is not relevant for the problem of false positives and the associated workload. In other applications, on the other hand, efficiency is more important and operating characteristics as positive predictive value will have more weight.

- *Relationship in the data*: some models are less flexible in the sense that they are only applicable to a smaller range of examples. For example, linear regression can only generate linear functions and therefore would be badly suited when the relationship in the data is non-linear. Other methods, on the other hands, offer a wider range of applications[76].

Expedited evidence: the time element

Being able to produce evidence when it is needed for decision-making or, in case of automation, to be synchronised with the business processes is a fundamental element to consider: not respecting this principle could cause the analysis results not to be useful and consequently to have no added value.

An example would be the detection of potential safety signals associated with newly marketed therapies: a perfect algorithm that identifies these concerns years later compared to more traditional methods would not be useful.

A key factor in influencing the speed is the availability of up-to-date data; this is mainly in control of the data provider, even if there are initiatives that can support it such as the availability of an IT infrastructure that allows fast access or exchange to the data. The elements considered below focus on how the speed of providing evidence is influenced by analytics:

Features of the data (or how data are prepared): data are usually messy. The inputs tend to be measured on very different scales and to be a mixture of quantitative, binary, and categorical variables, the latter often with different levels. Distributions of predictor and response variables are often long-tailed and highly skewed with outliers and there are generally missing values. Some models are able to deal with these characteristics better, other less and require more time for data transformation, cleaning and checks (see data manipulation chapter).

Computational scalability: data sets are often very large in terms of the number of observations and the number of variables measured and some models are more computationally intensive than others.

Internal feature selection: some models select the relevant variable as an integral part of the procedure. They are thereby resistant, if not completely immune, to the inclusion of many irrelevant predictor variables[76], others require more time to investigate which input variable to include.

---

[75] Dhasarathy A, Jain S, Khan N. When governments turn to AI: Algorithms, trade-offs, and trust. McKinsey and Company 2019 Feb.

[76] Hastie T, Tibshirani R, Friedman J. Springer 2016. The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition

Easiness to implement / simplicity: some models are quite easy to implement; others might have better performance (i.e. deep learning) but require more expertise and time to fine tune all the parameters.

Transparent evidence: the trust element

Results of AI and any advanced analytics models are used by people who usually are not the ones involved in producing them. Typically, when the final user is not the evidence generator, there tends to be questions around trust in the way the results have been produced[77]. Moreover, any algorithm developed is likely to vary with the nature of the regulatory decision, the strength of evidence required and, in particular, with the likelihood of deliberate falsification of data or analyses. When falsification is a possibility the question of whether the same algorithms as used by regulators to detect falsification will be available to the perpetrators and how this could be countered. Much of the discussion that follows will make the assumption, which is often true, that drug developers and regulators share a common aim to improve public health and act accordingly. However, in some areas, and particularly those where substantial economic repercussions may be associated with regulatory decisions, additional factors arise. These areas should be explicitly identified.

Trust can be enhanced by good governance in the production and handling of the evidence generated. Such an approach includes the following elements:

Interpretability (or explainability): providing a *qualitative understanding* of the possible causal logic and the relationship the model infers between input and predicted variables is becoming more and more important. This might be crucial in ensuring the models *make decisions for the right reasons*.

Also, the capability to *visualise the relationships* between variables and *communicate* them to a non-technical audience plays an important role in ensuring stakeholders and users understand the how and why of the results and that the results of the models are acted upon.

All of this in turn promotes confidence in the model: winning trust that the model will perform well with respect to the real objectives and scenarios.

But the benefits of interpreting machine learning models expand beyond troubleshooting and fixing errors. In some cases, they can help shed light on previously unknown aspects, providing insights into correlations that were not known allowing identification of new properties or generating new hypotheses that then could be tested.

The different available algorithms offer a wide range of spectrum along these qualities

Reproducibility: making reproduction of the results possible using the same or similar data sources. With machine learning algorithms reproducibility is more complex than with more traditional models and access to the data used in the analysis and knowledge about the type of design or model used might not be sufficient anymore.

The following will be required:

Access to the data used in the analysis.

Access to the analytics code: just 6% of machine learning papers share the algorithm code[78]. Reluctance to publish the code derives from the desire to keep ahead of competition, for commercial

[77] Schneeweiss S, Glynn R. Real-World Data Analytics Fit for Regulatory Decision-Making. American Journal of Law & Medicine, 44 (2017): 197-217

[78] Gundersen OE, Kjensmo S. State of the Art: Reproducibility in Artificial Intelligence. Associations for the Advancement of Artificial Intelligence 2018

interests but also because it often requires a significant additional effort to create publishable quality code. Moreover, it should be considered that sometimes the code will run only on specific software.

Access to the pre-processing code (see also data manipulation chapter): the codes provided should be self-containing, in practice there is often a separate code for pre-processing the data to input into the analysis. Without this code the only way to run the analysis would be to have the data from a certain domain all structured similarly, for instance having all data used in observational research (electronic health record, registries) in a common data model.

> This data handling code includes many decisions: if this code is not available and the specifics of those choices are not explicitly stated, a researcher trying to reproduce an analysis might obtain different results even when working on the same data source. An example in the pharmacoepidemiology analyses regards on whether the follow-up starts on the day of exposure or on the following day. Even such a simple choice could have a substantial impact: in the example of benzodiazepines and gastrointestinal bleeds, allowing the exposure to start the same day of the outcome created a spurious association, since it is more likely that benzodiazepines were prescribed for treatment of gastrointestinal bleeds[79].

Systematically capturing and reporting all the metadata associated with the final algorithm such as hyper-parameters, values of random initialisation, etc. Results might be very sensitive to those choices and it is recommended to create a guide specifying how and what should be reported.

> These considerations about reproducibility have a prominent role in the machine learning world, a hyper-competitive world where new algorithms and models are often evaluated using their performance on the test set. There is little reason for researchers to write or submit papers that propose methods with inferior test performance.
>
> As already mentioned in the methodological section, the test data should be held until the model has been developed using the training data so as to provide an unbiased estimate of the model's actual performance. However, using the test data in the training process has the advantage to increase the performance, even if this increase is due to overfitting. Some authors investigated this by creating a new test data for some published algorithms and found that across a wide range of different neural network models there was a significant drop in accuracy (4%−15%) from the published test set to the new test set. However, the relative ranking of each model's performance remained fairly stable[80].

Debugging: in some regulated industries, like finance and healthcare, it is required to audit the decision process and ensure it is not discriminatory or fraudulent.

Robustness and sensitivity analysis: in any analysis there tend to be some uncertainty regarding assumptions made and analytical choices. It would then be reassuring to test how robust the results produced are with slightly different choices, for instance changing the design or operationalise the variable of interest in different ways. This is even more important with more advanced machine learning algorithms that tend to be sensitive to the choices of the parameters used.

Continuous validation: constantly learning from new data offers opportunities for machine learning algorithms to keep improving their results. However, this opportunity could be challenging for

[79] Schneeweiss S, Glynn R. Real-World Data Analytics Fit for Regulatory Decision-Making. American Journal of Law & Medicine, 44 (2017): 197-217

[80] Chang O. Seven Myths in Machine Learning Research. 2019 Feb.

regulators that are used to dealing with more static clinical decision support tools; how could any deterioration in performance of the algorithm over time be detected?

Reconciling the need for dynamic solutions with the need for continuous validation could then be complex.

Moreover, algorithms results can influence decisions and this behavioural change was not present in the data used to train the initial algorithm: this could inadvertently cause the performance of algorithms to change over time. All models depend on using the past to predict the future and, over time, changes occur that invalidate the use of historical data to predict outcomes, either because of the natural evolution of the system studied (e.g. evolving health status) or as an intentional response to previous model results.

Trade-offs

The above discussion describes a number of characteristics that should be considered for regulatory applicability; some of them can be addressed with governance activities (for instance providing specific guidelines regarding data quality, how to publish and communicate results) or with infrastructure (for instance having enough computing power or the availability of specific software to enhance the speed of analysis or to run more advanced algorithms). Other characteristics, on the other hand, are inherent to the algorithm itself and can be mainly controlled by choosing an algorithm versus another.

In the latter case the trade-offs regarding the choices available are even more apparent: for instance, one algorithm can ensure speed and simplicity but at the expenses of being less accurate. This will be illustrated considering the trade-off between two of the most desired characteristics when selecting what model to use: prediction power (usually the most important requirement when the aim is to predict) and interpretability (critical when the aim is inference).

In the mid-1990s, a national wide effort was launched to create algorithms to predict for pneumonia patients, who should be admitted to hospital and who should be treated as outpatient. The goal was to send patients with a low risk for complications for outpatient treatment, preserving hospital beds and the attention of medical staff for those most at risk[81].

Initial findings indicated that neural networks were more accurate than classical statistical methods. However, doctors wanted to understand the 'thinking' behind this algorithm, so statisticians created 'decision rules' from the results of regression analysis. The examination of the results showed that both models (the neural networks and the regression) had inferred that pneumonia patients with asthma have a lower risk of dying, and therefore should not be admitted to hospital.

This outcome is counterintuitive, but it was the result of the models being successful in doing what they are designed to do: *discover and reflect a true pattern in the data*. Asthma patients with pneumonia usually were admitted not only to the hospital but directly to the ICU, treated aggressively, and survived.

The hospital anecdote highlights the practical value and importance of interpretability. Both models had learned that asthma is associated with lower risk, but the neural network was not easy to interpret, and the fact that medical attention confounded the relationship became difficult to diagnose. Only by interpreting the model was a crucial problem discovered and avoided.

[81] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1721-1730 (2015)

Even governments are starting to show concern about the increasing influence of impenetrable algorithms. The European Union recently proposed to establish a "right to explanation," which allows citizens to demand transparency for algorithmic decisions[82] and heavily penalises companies who cannot provide an explanation and record as to how a decision has been reached (whether by a human or computer).

Figure 20 shows a qualitative representation of the trade-off between model interpretability and model accuracy. On the top left, models that tend to prize explanation over accuracy, at the other end of the spectrum, models where accuracy is prized more.

The Figure 20 serves for illustrative purposes, the actual level of accuracy and interpretability of a model depends on the actual context: how the model is implemented and especially on the data used and the relationship between the variables in the data. For example, by adding more covariates one can make linear regression models more accurate but also less interpretable; tree models can be easier to interpret then regression model in case of discrete variables, etc.

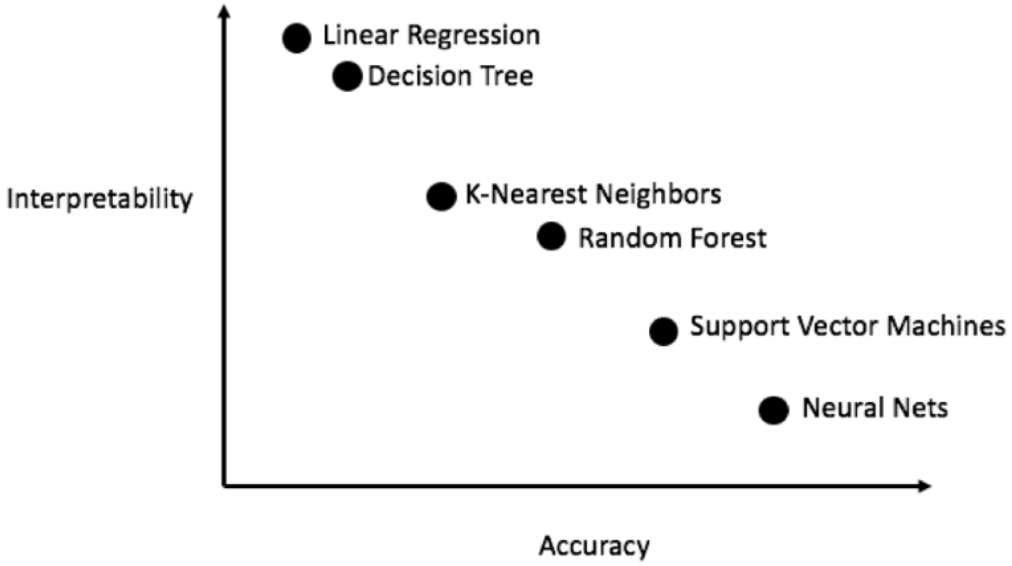


Figure 20: What vs why: a representation of the trade-off between interpretability and accuracy

Source: Interpreting Machine Learning Models https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9.

This way of looking at the different models helps to explain why linear regression has been so popular: because it is the gold standard in interpretability. In fact, linear regression output clearly describes both the magnitude and the direction of each predictor.

Figure provides a conceptual framework of the balance between interpretability and accuracy and seems to suggest that modern machine learning offers a choice among what and why: is it preferred to know what will happen with high accuracy, or why something will happen? The answer is contextual and influenced by the main aim of the model: the "why" helps in strategising, generating confidence

[82] Metz, C. Artificial Intelligence Is Setting Up the Internet for a Huge Clash with Europe. Wired.com (2016) and Goodman & Flaxman, 2016

and better know the limitations of the model; the "what" helps acting with higher precision in the immediate future.

In some areas where trust is important, solutions to improve interpretability have been prioritised: for instance, the Uppsala Monitoring Center has developed a new predictive model called vigiRank that considers multiple dimensions, in addition to disproportionate reporting patterns, to prioritise emerging safety signals[83]. To facilitate its interpretability, not only the final score of the algorithm is displayed, but the value of each dimension considered (disproportionality reporting, completeness, recency, geographic spread and availability of case narratives) is shown to make the assessor understand what drive a specific score.

In other areas the opposite is true. For instance, if machine learning algorithms are used to de-identify clinical narratives, the interest lies more in having an accurate algorithm. These results will not support complicated decisions, but it is highly sensitive that all the possible identification information is hidden; the trust derives more from the high accuracy of the model used[84].

Moreover, the answer to the same question can change over time: in fact, organisations can also consider moving to more complex algorithms once the user base becomes more familiar with and trust is built in the more explainable models[85].

Finally, rather than being static, the accuracy vs. interpretability trade-off is now seen as a frontier that is being pushed outwards; allowing more interpretability without losing accuracy (or more accuracy without losing interpretability). For the data science community (academic institutions, government agencies, and tech companies) developing more interpretable models or models that are more transparent and open to investigation is becoming as important as developing more accurate models. One approach to be mentioned in this area is LIME (Local Interpretable Model-Agnostic Explanations)[86], which allows building any model, and then using linear approximation to explain specific predictions.

These last examples have described the rise of interpretability, how it is becoming the core of the choice of a model and as designing algorithms from the start with interpretability in mind is becoming critical. As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic[87]. And if the user of these models does not understand, the overall performance of the system might be lower than a lower-performing model that the user is able to understand. Promoting trust in good algorithms,and the ability to detect when algorithms take decisions not for the right reason, is critical.

On the other hand, careful considerations should also be given about when giving up predictive power; the desire for transparency should also be justified and not being simply a concession to biases against new methods (in this regard, it is also notable that many aspects of the practice of medicine are unexplained, such as prescription of a drug without a known mechanism of action[62]). As a concrete example, it is important to balance the goal of building trust with doctors by developing transparent models with the longer-term goal of improving health care.

---

[83] Caster O, Sandberg L, Bergvall T, Watson S, Norén GN. Pharmacoepidemiol Drug Saf. 2017;26(8):1006-1010

[84] Presentation by Ellenius J. Opportunities and barriers for machine learning to improve patient outcomes. ICPE Prague 2018

[85] Dhasarathy A, Jain S, Khan N. When governments turn to AI: Algorithms, trade-offs, and trust. McKinsey and Company 2019 Feb.

[86] Local Interpretable Model-Agnostic Explanations (LIME): An Introduction.

[87] Lipton CZ. The mythos of model interpretability. Cornell University 2016 Jun.

4.6. Regulatory implications

The regulatory implications of AI can be divided into two broad groups: i) interactions with the stakeholders and ii) internal developments.

4.6.1. Interactions with stakeholders

The approach to regulation of AI to protect public health should be informed by the opposing needs to encourage useful innovations and manage risks.

The starting point would be to consider whether the existing regulations and governance already adequately address this trade-off, or whether there is a need for them to be adapted to the specificity of more advanced machine learning algorithms. The key aspect to consider is how to improve the trade-off, how to lower costs and barriers to innovation without adversely impacting safety[88].

In this regard, having an open dialogue with the applicant for an authorisation would be helpful to go through the planned analysis so that the potential for bias is transparent. When a MAH is considering using more advance machine learning algorithms to support a marketing authorisation, a scientific advice procedure could be used, and this might be extended to validating the methods and advising on its acceptability for regulatory use considering the framework and characteristics described previously. Such interaction would also help to develop and refine over time a regulatory knowledge base and familiarise with practical applications of AI and should leverage the expertise already presented in the Biostatistics and Modelling and Simulation Working Party and Pharmacogenomics Working Party.

Considering the area of research and publications of machine learning algorithms, this has become one that emphasizes winning, focusing on demonstrating how a new method beats previous ones and not *on the process for developing insights, understandings, and based on empirical rigour*. This may have led to a negative effect on the research[89]. While the research environment has started to acknowledge this, influencing this process and highlighting the key requirements and methodological rigour required by regulatory applicability would help.

4.6.2. Internal

As regulator, a stepwise approach to use of AI may be needed. Considering the two areas in which the AI opportunities have been classified, priority might be given to select regulatory activities where AI could bring added value to increase efficiency, as this objective could be more acceptable than to support decision-making, which may initially raise concerns about trust and interferences with expert assessment. One aim of this report is indeed to develop a framework to help this process; both by setting principles that an algorithm has to respect, but also by empowering assessors highlighting areas where attention should be paid to critically challenge the findings of such algorithms in order to be confident that the derived evidence is trustworthy. Improved human-technology interactions should not raise concerns and being seen as a positive development.

To facilitate evaluation and possible adoption of AI initiatives, the following areas should be considered:

[88] Preparing for the future of artificial intelligence. Executive Office of the President National Science and Technology Council Committee on Technology. 2016 Oct.
[89] Sculley D, Snoek J, Rahimi A, Wiltschko A. Winner's curse? Workshop track - ICLR 2018

Data sources: to identify and evaluate relevant data sources of sufficient quality and ensure their continuous availability.

Piloting: start small, with methods which are sufficiently transparent, incremental to something already in place or small in scope, and which may bring immediate value to the system. The aim of the piloting is to develop internal competencies and enrich knowledge to serve as input for more routine and complex applications.

Fostering internal capability: not only building expertise in data science skills but also creating an environment where subject matter experts can work together with data scientists and creating space for experimentation.

Creating an expert working group: once expertise has been gained through piloting, small initiatives and collaboration, a dedicated group to provide advice and scale these initiatives will be beneficial. This expert group should also collaborate and gather input from external experts.

Addressing regulatory acceptability: agree and create guidelines on which level of validation, reproducibility and trustworthiness of evidence is acceptable according to the regulatory purpose of application of the AI algorithm.

Monitoring of developments: the interest and the number of researchers and initiatives in this field are growing considerably, creating opportunities for collaborations. Engagement with the research community, the European Commission and other regulators to identify new directions, shaping priorities and to respond to existing opportunities.

Fostering communication: one barrier to adoption and innovation of AI is probably not the data science itself but the communication[90]. When the work is not understood, it is not used and serves no purposes. Providing clear and good examples of applications in an easy to understand way, will help involving people, creating a community, increase the organisational capacity to use and consume data as part of decision making and increased support from senior management.

4.7. Conclusions

There are two stories about AI: one describes all the opportunities amplified by the golden promise of AI companies; the other is that AI raises two major concerns: bias and vulnerability to abuse. This report focused on demystifying both stories, being more realistic on the opportunities but also describing a clear framework to reduce the challenges.

In terms of opportunities, the greatest potential for the newest AI methods reside in two broad categories. One category is to create value in use cases in which more established analytical techniques such as regression techniques can already be used, but where deep learning techniques could provide higher performance or generate additional insights and applications. The other category is to be able to deal with problems that could not be automated until few years ago such as images, speech and video.

In terms of challenges, healthcare in general and its regulatory field may have adopted less in the area of AI due to having a very low tolerance for errors, and therefore not feeling the need to invest in staff that have both domain-knowledge in health-care as well as the sophisticated mathematical modelling and data science skillset. This may not have encouraged a receptive attitude to new technologies.

[90] Livni E. Storytellers make the most influential scientific researchers. Quartz.com 2016 Dec.

Another challenge is about the skillset; training activities and collaboration with academia and experts to increase the expertise in the network and ensure critical capacity is reached will need to be planned. Moreover, from an organisational aspect, co-operation of data scientist and subject-matter experts (pharmaceutical and clinical experts) should be nurtured as it is the key element to develop models appropriate to the question off interest benefiting from expertise that cannot be in only one resource.

Another obstacle not dealt in this report relates to the perceived impact on the workforce. The common view is to see AI competing versus humans and this creates resistance. However, the role of human will always be important, from creating and testing the models where many choices require intuition from analysts and subject matter experts, to working with the model results and monitoring it. Placing the human at the centre of AI applications and making human working together with AI (what is referred as Augmented Intelligence) can lead to substantially better decisions than the human or the machine alone. And this is reflected even in research that shows how the impact of AI is neural in terms of the number of resources, but significant in terms of the skills required.

A framework of principles to assess the regulatory applicability of AI algorithms has been proposed; following it might provide a more systematic evaluation of developments in the area. In particular, to evaluate its ability to support regulatory decision-making it is important to:

Have access to high quality data: if the answers sought are not included in the information content of the data, models or algorithms will not be able to uncover them no matter how elaborate they are. It is therefore critical to develop a strategy for collecting and acquiring the relevant data.

> Moreover, sometimes data might exist but is stored in silos. The growth and value added of big data and analytics have been based on combining different kinds of data, which enhances its information content. Consequently, encouraging data sharing and integration are essential to ensure the success of AI models.
>
> Finally, data should always be evaluated for potential bias:

Derive valid and reliable insights and answers: as AI algorithms cannot uncover information which is not present in the data, they also cannot resolve the usual methodology flaws when working with observational data about *causal inference* and *bias*.

> The main theme is that even the more advanced methods will work by building on the old statistical lessons and principles, not by ignoring or deviating from them.

Provide timely evidence: insights and answer should be produced when needed, synchronised with the decision-making and/or business process.

Characterise advanced algorithms accurately: the additional challenge of explaining in human terms results from large and complex models, why a certain decision was reached, is particularly relevant as regulators often want rules and choice criteria to be clearly explainable.

> Many situations arise when real world objectives are difficult to encode as simple algorithmic functions. Typically, problems such as ethics and legality cannot be directly included in an algorithm and optimized. Human interpretations of the analysis results serve those objectives that we deem important but struggle to model formally.

The main message is that independently of how complex an algorithm is, we should always go back to basic and critically examine the data used, the methodology and how the algorithm got the results. And this is true for any kind of analysis and should be applied not only for AI: why would the use of Excel (crammed with complex and invisible formulae) not be an issue to support decision making, but the use of AI be one?

Finally, in this report the distinction between prediction and causal inference was crucial in defining AI and its applications. However, prediction, or mapping observed input to output, does not cover all realms of intelligence (and it could be argued that it barely covers either of them). A more challenging sign of intelligence is the ability to answer causal inference questions and to reason counterfactually. No type/form of AI will be worthy of the name without this ability, and a more critical approach to the results produced might benefit the development of AI applications and scope.

4.8. Recommendations

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
1	Data availability and quality	Data is an enabler for AI, need to ensure access to quality and representative data (even with advanced AI algorithms the concept of 'garbage in – garbage out' still applies)	Increase the availability of and access to high quality data. Establish a data quality framework to assess, compare and understand the quality of the data and establish quality attributes considering the different requirements for each type of data. Examine and report the quality of the data used to train the model in term of consistency, completeness, accuracy and representativeness. Incentivise sharing of the data and new paradigm of data ownership (e.g. federation of data). Establish horizon scanning process to identify important data gaps and develop strategies to fill them.	Increased reporting of critical assessment of the quality of the data used in the algorithms
2	Timely evidence	Produce evidence synchronised with the decision-making or, in case of automation, with the business processes	Implement an infrastructure that allows fast access or exchange to data considering the different requirements from the different data sources (internal versus external, aggregated versus patient level). Increase computational scalability to efficiently and timely analyse a very large number of observations and variables. Choose algorithms and models based also on simplicity and easiness of implementation when timely evidence is critical (the validation elements, described below, must still apply).	Increased availability of timely evidence

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
			Support the development of analytical software, preferably open source, to accelerate and standardise the evidence generation.	
3	Choice of the algorithm/model	"Make your model as simple as possible, but not simpler"	Characterise the strengths and limitations of machine learning algorithms. Ensure that the following elements are considered when choosing the algorithm / model: The question of interest. The ultimate aim, whether prediction, inference, or a combination of the two. The complexity of the data (e.g. non linearity, dimensionality, interactions between variables). Considerations about the data available, timely evidence, validation and transparency as described in the rest of the recommendations. Start simple, the most sophisticated techniques are not always the most appropriate. Support continuous update and development of guidance.	Increased reporting of reasoning about the algorithm/model choice Guidance document
4	Validity of the analysis	Develop a framework to validate machine learning algorithms used to support regulatory decision making	Develop robust validation procedures. Support the development and continuous update of guidance on methodological elements to consider in machine learning algorithms. Support the critical assessment and validation of the algorithm results; this assessment become even more	Increased reporting of critical assessment of the methodological elements in machine learning algorithms used to

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
			important as millions of variables and observations are included in the model and more precise results of unknown accuracy are generated. Require that the model results are obtained on data that are either truly independent validation data sets, from different populations or periods that played no part in model development, or using cross validation techniques. Require the reporting of all the relevant accuracy measures (e.g. sensitivity, specificity) for the specific situation in which the model will be used. Strongly encourage the use of negative controls to assess the model performance. Develop a clear framework and process for the validation of models including: Consideration about a qualification procedure for machine learning algorithms, where privacy of the algorithm content is retained. Encouraging MAHs to interact early in the development process when using advanced machine learning models so that the potential bias is transparent from the beginning.	support regulatory decision making Guidance document to support the validation of analyses based on machine learning algorithms
5	Transparency and trust	Promote transparency and audit of machine learning algorithms; appropriate governance in the production and handling of the evidence	Require reporting clear justification of database choice, study design and subsequent protocol changes. Enhance the interpretability of the model:	Public availability of protocols and analysis plans for all studies submitted for regulatory approval

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
		generated should be established	Encourage to consider the potential trade-off between interpretability and predictive power from the start of the model development. Encourage to assess and provide a qualitative understanding of the possible causal logic and the relationship the model infers. Encourage the comparison of the results of more advanced models with simpler ones to assess the increase in performance versus the potential lost in interpretability. Support the development of guidelines to increase reproducibility outlining how and when to provide access to i) the data used; ii) the pre-processing code; iii) the analysis code; iv) the metadata associated with the final algorithm. Encourage to assess the robustness of the results using sensitivity analyses based on different choices of the parameters used in the model. Assess the need for continuous validation and monitor of the performance of the model.	Establish guidelines for when and how to provide access to the data and code used for the analysis
6	AI algorithms used to support automation	Define in advance, considering the data source used and the final objective of the application, realistic error rates, how to best integrate the application in the business process and how to measure the benefits	Support the same critical assessment as reported in the 'validity of the analysis' topic mentioned above. Identify processes (or specific parts of a process) where simple improvements / optimisations can lead to significant returns (i.e. solve the right problems, not the most stimulating ones).	Successful implementation of AI algorithms for automation with benefits measured

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
			Define in advance an acceptable error rate based on the specific situation (e.g. when sensitivity or specificity is more important). Consider how the automation will fit in the business process. Identify opportunities where a semi-automatic approach would provide better returns (for part of the processes more difficult to fully automate). Consider opportunities to establish a process where human feedback on the results of the AI algorithm are used to improve the performance. Establish the need for a continuous monitoring of the algorithm performance. Measure the return of the investment to prove the benefits of the solution applied.	Increased consistency in business processes using AI algorithms
7	Opportunities	Identify opportunities where application of advanced machine learning algorithms can contribute to better performance compared to currently used analyses	Explore natural language processing techniques to extract structured information from various data sources that can be used for i) queries and searches; ii) monitoring over time and iii) as input to advanced analytics models. Explore novel analytical methods to have a faster and more accurate i) diagnosis; ii) prediction of future outcomes; iii) identification of adverse drug reactions.	Increased efficiency and value of the use of unstructured data Improved performance of methods to support the development, evaluation and monitoring of medicinal products

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
8	Communication	Demystify the use of advanced methods and provide confidence to the end users of how advanced models can help	Demystify AI and its consequences: identify real opportunities and limits. Provide a clear and simple understanding of how the models work, their results and how they can be used. Communication and visualisation training for data scientists.	More balanced and informed discussion about AI potentials and limitations
9	Collaboration	Engage with research community and on-going initiatives	Proactive engagement with international organisation, other regulators, European Commission and academia. Leverage the outcome of EU projects. Actively promote a collaborative approach to research. Monitor the development of innovative methods with applications to medicines. Demonstrate a commitment to open science tools and algorithms to facilitate accessibility of research and analyses. Influence the research community to focus on developing insights and understandings using empirical rigour instead of focusing on demonstrating how new methods beats previous ones. Identify, encourage and endorse best practices.	Active participation in external initiatives
10	Organisation	Promote centre of excellences, collaboration between data scientists and subject matter experts, and a culture of experimentation	Leverage the expertise already available in the Working Parties (Biostatistics and Modelling & Simulation). Explore case studies and promote pilots to investigate the utility and benefits of new machine learning methods.	Establish Analytics Centre of Excellences

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
			Foster the collaboration between data scientists and subject matter experts. Promote a culture of experimentation to gain hands-on insights on the novel models. Establish dedicated Analytics Centre of Excellences in the EU regulatory network to gain insights in the use of innovative methods and experiment and develop prototypes in collaboration with relevant business areas: Promote collaboration between the different Analytics Centre of Excellences. Define their role in the regulatory process. Require strong collaboration with subject matter experts to exchange and develop ideas. Create an AI Advisory Group of experts including experts and links with academia and research centres to: Explore the applicability of novel analytics methodologies related to the development, evaluation and monitoring of medicinal products. Provide input in the development and update of guidelines. Encourage crowdsourcing projects sharing non-sensitive and anonymised datasets to the external world to accelerate the improvement of the performance of the models currently used. Promote change in culture in:	Establish an AI Advisory Group

#	Topic	Core recommendation	Reinforcing actions	Evaluation criteria
			Perceiving tasks as prediction problems. Perceive AI as possibility to enhance human capabilities (it is not AI versus humans, but humans augmented by AI vs human working without AI). Support moving to more complex algorithms once the user base becomes more familiar with and trust is built on the use of advanced methods.	
11	Skills and resources	Develop the capacity to allow a critical appraisal of studies done with advance model and/or to perform these studies	Establish a curricula or profile of skills needed Identify internal resources with existing skills or potential to be trained Provide training to be able to assess and/or use these rapidly evolving analytics. Hire experts in AI and related fields. Promote a culture of experimentation to gain hands-on insights on the novel models. Develop competencies with collaborations.	Increase regulatory capacity for analysis and assessment on the use of advanced algorithms

5. Annex A

5.1. Description of the most well-known Standardisation Development Organisations (SDOs)

ISO/CEN

According to the ISO website: "ISO is an independent, non-governmental international organization with a membership of 161 national standards bodies. Through its members, it brings together experts to share knowledge and develop voluntary, consensus-based, market relevant International Standards that support innovation and provide solutions to global challenges".

Of particular interest would be:

ISO/IEC JTC 1 – Information Technology

Scope: Standardization in the field of information technology.

ISO/IEC JTC 1/SC 42 - Artificial intelligence (which include also a WG on Big Data)

a. SC 42 Scope: Standardization in the area of Artificial Intelligence
    - Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence
    - Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications
b. SC 42/WG 2 Scope: Standardization in the area of Big Data
    - Serve as the focus and proponent for JTC 1's standardization program on Big Data.

ISO/TC 215 - Health informatics

Scope: Standardization in the field of health informatics, to facilitate capture, interchange and use of health-related data, information, and knowledge to support and enable all aspects of the health system.

HL7

Founded in 1987, Health Level Seven International (HL7) is a not-for-profit, ANSI-accredited standards developing organisation supported by more than 1600 members from over 50 countries representing healthcare providers, government stakeholders, payers, regulators, pharmaceutical companies, vendors/suppliers, and consulting firms. HL7 is dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services. There are more than 50 Work Groups participating in various standardisation projects and initiatives with a number of them directly or indirectly related to the business domains referred above (see http://www.hl7.org/Special/committees/index.cfm?ref=nav).

The HL7 Biomedical Research and Regulation (BR&R) WG areas of interest include clinical and translational research (both regulated and non-regulated) and the subsequent regulatory submissions and information exchanges to bring new products to market and to ensure safe use throughout the product lifecycle. The group is currently working on a number of standardisation projects including the FHIR resources for ISO IDMP, CDISC Lab Semantics in FHIR, mapping work related to Common Data

Model Harmonisation (CDMH) - FHIR Implementation Guide, updating the BRIDG model (a standard developed by HL7, CDISC and ISO SDOs), etc.

The HL7 Clinical Genomics WG is focusing at the personalisation (differences in individual's genome) of the genomic data and the linking to relevant clinical information. It collects, review, develop and document clinical genomics use cases in order to determine what data needs to be exchanged, as well as, reviewing existing genomics standards formats such as BSML (Bioinformatics Sequence Markup Language), MAGE-ML (Microarray and GeneExpression Markup Language), LSID (Life Science Identifier) and others.

The developments in genomics and precision medicine together with the diversity of collecting, sharing, coding and exchanging genomic information from various sources has resulted in different terminologies and infrastructures that limit semantic interoperability and data analysis. The Global Alliance for Genomics and Health (GA4GH) has been working by building and refining an API and data model for the exchange of full sequence genomic information across multiple research organisations and platforms. The HL7 FHIR technology is another attractive approach. It is based on a set of modular components called resources; it is easier to implement, providing a very functional framework to initiate interoperable clinical genetics data standardisation.

Sync for Genes is an initiative aiming at developing a standardised method for sharing genomic data, enabling integration of clinical genomics information to an electronic health record (EHR) for better patient care via point-of-care apps (e.g., SMART on FHIR Genomics, patient-facing apps, as well as clinical research and analysis). Sequence is a new FHIR resource that will be used to hold clinically relevant sequence data in a manner that is both efficient and versatile integrating new and as yet undefined types of genomic and other -omics data that will soon be commonly entered into health records for clinical use and decision making. Sequence can be extended to address complex cases and profiles, can be able to support a large set of clinical use cases and is thus positioned to address all emergent -omics use cases, including Next-Generation Sequencing (NGS).

The HL7 RDF for Semantic Interoperability WG is jointly collaborating with the W3C Healthcare and Life Sciences group on Clinical Observations Interoperability to facilitate the use of RDF as a common semantic foundation for healthcare information interoperability. One important feature of RDF is that it captures information content independent of syntax or data formats. This ability enables RDF to act as a universal information representation and use existing data formats, but each one can have an RDF equivalent that unambiguously captures its intended information content. Instead of replacing or creating new standards RDF can be used to achieve a cohesive mesh of standards that can be used together, as though they constitute a single comprehensive standard, aka, standardising the standards. RDF can play an important role in semantic interoperability not only on standardisation aspects (let’s say exchanging information using the same the data models and vocabularies) but also on translation aspects (translating between data models and vocabularies).

CDISC

CDISC is a global, non-profit charitable organization that develops data standards to streamline clinical research and enable connections to healthcare, empowering the valuable information offered by patients participating in research studies around the world. CDISC has developed several kinds of standards serving different purposes.

ICH

As we read from its website: "The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) is unique in bringing together the regulatory authorities and pharmaceutical industry to discuss scientific and technical aspects of drug registration. Since its inception in 1990, ICH has gradually evolved, to respond to the increasingly global face of drug development. ICH's mission is to achieve greater harmonisation worldwide to ensure that safe, effective, and high-quality medicines are developed and registered in the most resource-efficient manner".

GS1

GS1 Healthcare is a neutral and open community bringing together all related healthcare stakeholders to lead the successful development and implementation of global GS1 standards enhancing patient safety, operational and supply chain efficiencies. GS1 standards create a common foundation for a business by uniquely identifying, accurately capturing and automatically sharing vital information about products, locations and assets. GS1 standards are now present across many sectors such as healthcare, fresh foods and foodservice.

IEEE

IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.

IEEE's mission statement: IEEE's core purpose is to foster technological innovation and excellence for the benefit of humanity.

IEEE's vision statements: IEEE will be essential to the global technical community and to technical professionals everywhere and be universally recognised for the contributions of technology and of technical professionals in improving global conditions.

Of particular interest would be:

IEEE P7000 working group - Model Process for Addressing Ethical Concerns During System Design

> Scope: The standard establishes a process model by which engineers and technologists can address ethical consideration throughout the various stages of system initiation, analysis and design. Expected process requirements include management and engineering view of new IT product development, computer ethics and IT system design, value-sensitive design, and, stakeholder involvement in ethical IT system design.

IEEE Big Data Initiative :

> Scope: Big data is much more than just data bits and bytes on one side and processing on the other. It entails collecting, storing, processing, and analysing immense quantities of data that is diverse in structure in order to produce insights that are actionable and value-added. Vast amounts of data of various types are being generated at increasing rates. Determining how to utilize this data strategically and efficiently is the goal of technologies associated with the Big Data initiative.
>
> Merely collecting and storing data is not the sole objective of Big Data; rather, enhancement of businesses or societies drives the technologies of Big Data. For example, successful big data solutions can provide targeted marketing, identify new markets, or improve customer service through analysis of customer data, social media, or search engine data. Examination of industrial

sensor data or business process data can enhance production, aid in proactive improvements to processes, or optimize supply chain systems. As a final example, society can benefit from big data analytics through intelligent healthcare monitoring, cybersecurity efforts, and smart cities data manipulation.

There are multiple challenges associated with Big Data, including:

- Recognition of useful versus irrelevant data,
- Collection of distributed data,
- Accuracy, completeness, and timeliness of data,
- Efficient storage and transfer,
- Privacy and security of data,
- Fault tolerance,
- Scalability and economic impact of implementation,
- Intelligent analysis,
- Insightful and flexible presentation.

The IEEE Future Directions Big Data Initiative strives to aggregate information about the various endeavours occurring worldwide in order to provide a community of professionals in industry, academia, and government working to solve the challenges associated with Big Data. Through various outlets, participants in the Big Data Technical Community can learn and collaborate on the multi-faceted Big Data initiative that has applications in many industries and markets. Members of the community have access to extensive resources including publications, videos, articles, interviews, webinars, newsletters, workshops, and conferences.

IEEE Future Directions Big Data Initiative also worked with IEEE Standards Association for establishing the Big Data Governance and Metadata Management to explore standard reference architecture for Big Data governance and metadata management that is scalable and can enable the Findability, Accessibility, Interoperability, and Reusability between heterogeneous datasets from various domains without worrying about data source and structure. The goal to enable data integration/mashup among heterogeneous datasets from diversified domain repositories and make data discoverable, accessible, and usable through a machine readable and actionable standard data infrastructure.

5.1.1. List of relevant standards

Research has been undertaken to identify all the standards developed by the main SDOs within the specific committees relevant for Big Data in the context of regulated medicines.

The descriptive text below is taken verbatim from the various SDOs websites:

a) Under the umbrella of ISO/IEC JTC 1/SC 42 and ISO/TC215:

1. ISO/IEC AWI TR 20547-1 [Under development] - Big data reference architecture - Part 1: Framework and application process.
2. ISO/IEC IS 20546:2019 [Published] - Big Data – Overview and vocabulary.
3. ISO/IEC TR 20547-2:2018 [Published]- Big data reference architecture -- Part 2: Use cases and derived requirements.

ISO/IEC TR 20547-2:2018 provides examples of big data use cases with application domains and technical considerations derived from the contributed use cases.

4. ISO/IEC DIS 20547-3 [Under development] - Big data reference architecture -- Part 3: Reference architecture.
5. ISO/IEC IS 20547-4 [Under development by ISO/IEC JTC 1/SC 27] – Big data reference architecture – Part 4: Security and privacy.
6. ISO/IEC TR 20547-5:2018 [Published]- Big data reference architecture -- Part 5: Standards roadmap.
7. ISO/IEC AWI 22989 [Under development] - Artificial Intelligence Concepts and Terminology.
8. ISO/IEC AWI 23053 [Under development] - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).

ISO/IEC TR 20547-5:2018 describes big data relevant standards, both in existence and under development, along with priorities for future big data standards development based on gap analysis.

9. ISO 11238:2018 [Published]- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of regulated information on substances.

This document provides an information model to define and identify substances within medicinal products or substances used for medicinal purposes, including dietary supplements, foods and cosmetics. The information model can be used in the human and veterinary domain since the principles are transferrable. Other standards and external terminological resources are referenced that are applicable to this document.

10. ISO 11239:2012 [Published]- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of regulated information on pharmaceutical dose forms, units of presentation, routes of administration and packaging.

ISO 11239:2012 specifies:

- the structures and relationships between the data elements required for the exchange of information, which uniquely and with certainty identify pharmaceutical dose forms, units of presentation, routes of administration and packaging items related to medicinal products;

- a mechanism for the association of translations of a single concept into different languages;

- a mechanism for the versioning of the concepts in order to track their evolution;

- rules to allow regional authorities to map existing regional terms to the terms created using ISO 11239:2012 in a harmonized and meaningful way.

11. ISO 11240:2012 [Published]- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of units of measurement.

ISO 11240:2012:

- specifies rules for the usage and coded representation of units of measurement for the purpose of exchanging information about quantitative medicinal product characteristics that require units of measurement (e.g. strength) in the human medicine domain;

- establishes requirements for units in order to provide traceability to international metrological standards;

- provides rules for the standardised and machine-readable documentation of quantitative composition and strength of medicinal products, specifically in the context of medicinal product identification;

- defines the requirements for the representation of units of measurement in coded form;

- provides structures and rules for mapping between different unit vocabularies and language translations to support the implementation of ISO 11240:2012, taking into account that existing systems, dictionaries and repositories use a variety of terms and codes for the representation of units;

The scope of ISO 11240:2012 is limited to the representation of units of measurement for data interchange between computer applications.

12. ISO 11615:2017 [Published]- Identification of medicinal products -- Data elements and structures for the unique identification and exchange of regulated medicinal product information.

ISO 11615:2017 establishes definitions and concepts and describes data elements and their structural relationships, which are required for the unique identification and the detailed description of Medicinal Products. Taken together, the standards listed in the Introduction define, characterise and uniquely identify regulated Medicinal Products for human use during their entire life cycle, i.e. from development to authorisation, post-marketing and renewal or withdrawal from the market, where applicable. Furthermore, to support successful information exchange in relation to the unique identification and characterisation of Medicinal Products, the use of other normative IDMP messaging standards is included, which are to be applied in the context of ISO 11615:2017.

13. ISO 11616:2017 [Published]- Identification of medicinal products -- Data elements and structures for unique identification and exchange of regulated pharmaceutical product information.

ISO 11616:2017 is intended to provide specific levels of information relevant to the identification of a Medicinal Product or group of Medicinal Products. It defines the data elements, structures and relationships between data elements that are required for the exchange of regulated information, in order to uniquely identify pharmaceutical products. This identification is to be applied throughout the product lifecycle to support pharmacovigilance, regulatory and other activities worldwide. In addition, ISO 11616:2017 is essential to ensure that pharmaceutical product information is assembled in a structured format with transmission between a diverse set of stakeholders for both regulatory and clinical (e.g. e-prescribing, clinical decision support) purposes. This ensures interoperability and compatibility for both the sender and the recipient. ISO 11616:2017 is not intended to be a scientific classification for pharmaceutical products. Rather, it is a formal association of particular data elements categorised in prescribed combinations and uniquely identified when levelling degrees of information are incomplete. This allows for Medicinal Products to be unequivocally identified on a global level. References to other normative IDMP and messaging standards for pharmaceutical product information are included in Clause 2, to be applied in the context of ISO 11616:2017. Medicinal products for veterinary use are out of scope of ISO 11616:2017.

14. ISO 12052:2017 [Published]- Digital imaging and communication in medicine (DICOM) including workflow and data management.

ISO 12052:2017, within the field of health informatics, addresses the exchange of digital images and information related to the production and management of those images, between both medical

imaging equipment and systems concerned with the management and communication of that information. ISO 12052:2017 facilitates interoperability of medical imaging equipment by specifying:

- for network communications, a set of protocols to be followed by devices claiming conformance to this document;

- the syntax and semantics of Commands and associated information which can be exchanged using these protocols;

- for media communication, a set of media storage services to be followed by devices claiming conformance to this document, as well as a File Format and a medical directory structure to facilitate access to the images and related information stored on interchange media;

- information that is to be supplied with an implementation for which conformance to this document is claimed.

ISO 12052:2017 pertains to the field of medical informatics. Within that field, it addresses the exchange of digital information between medical imaging equipment and other systems. Because such equipment may interoperate with other medical devices and information systems, the scope of this document needs to overlap with other areas of medical informatics. However, this document does not address the full breadth of this field. ISO 12052:2017 has been developed with an emphasis on diagnostic medical imaging as practiced in radiology, cardiology, pathology, dentistry, ophthalmology and related disciplines, and image-based therapies such as interventional radiology, radiotherapy and surgery. However, it is also applicable to a wide range of image and non-image related information exchanged in clinical, research, veterinary, and other medical environments. ISO 12052:2017 facilitates interoperability of systems claiming conformance in a multi-vendor environment, but does not, by itself, guarantee interoperability.

15. ISO/TR 12773-1:2009 [Published]- Business requirements for health summary records -- Part 1: Requirements.

ISO/TR 12773-1:2009 is based on a comprehensive review of a series of initiatives and implementations worldwide that for the purposes of this Technical Report are collectively called health summary records (HSRs). Project sponsors and/or authorities were contacted as needed to gather additional information and clarify questions or issues arising out of the review. ISO/TR 12773-1:2009 defines and describes HSRs in general as well as specific instances of HSRs and their most common use cases. It summarises the business requirements driving HSR development and the content that is common across HSRs, as well as issues associated with them. Finally, it recommends some future ISO/TC 215 activities to support international standardisation of HSRs. It is important to note that ISO/TR 12773-1:2009 focuses primarily on requirements that are specific (unique) to HSRs. It does not attempt to articulate, other than at a high level, requirements that are generally applicable to all health records or all electronic health records.

16. ISO/TR 12773-2:2009 [Published]- Business requirements for health summary records -- Part 2: Environmental scan.

ISO/TR 12773-2:2009 reviews a series of initiatives and implementations worldwide that for purposes of this Technical Report are collectively called health summary records (HSRs). It provides an environmental scan and descriptive information on HSR initiatives internationally, including "lessons learned". The environmental scan was completed by performing web searches and obtaining publicly available documentation on key projects. Project sponsors and/or authorities

were contacted as needed to gather additional information and clarify questions and issues arising out of the review.

17. ISO/TR 13054:2012 [Published]- Knowledge management of health information standards.

ISO/TR 13054:2012 describes a standards knowledge management (SKM) methodology and metadata to support the easy identification of the existence of a health informatics standard, its developmental status, and its associated Standards Development Organization (SDO). In particular, it describes a knowledge-based navigation methodology to enable rapid appreciation of the contextual roles and purposes of a standard, including the relationship between one standard and others, particularly in the same standards domain. ISO/TR 13054:2012 also gives information about the design of tools to support knowledge management of health informatics standards.

18. ISO/TS 13131:2014 [Published]- Telehealth services -- Quality planning guidelines.

ISO/TS 13131:2014 provides advice and recommendations on how to develop quality objectives and guidelines for telehealth services that use information and communications technologies (ICTs) to deliver healthcare over both long and short distances by using a risk management process. The following key requirements are considered when developing quality objectives and guidelines for telehealth services:

- management of telehealth quality processes by the healthcare organization;

- management of financial resources to support telehealth services;

- processes relating to people such as workforce planning, healthcare planning, and responsibilities;

- provision of infrastructure and facilities resources for telehealth services;

- management of information and technology resources used in telehealth services.

19. ISO 13606-1:2008 [Published]- Electronic health record communication -- Part 1: Reference model.

ISO 13606-1:2008 specifies the communication of part or all of the electronic health record (EHR) of a single identified subject of care between EHR systems, or between EHR systems and a centralized EHR data repository. It may also be used for EHR communication between an EHR system or repository and clinical applications or middleware components (such as decision support components) that need to access or provide EHR data, or as the representation of EHR data within a distributed (federated) record system. ISO 13606-1:2008 will predominantly be used to support the direct care given to identifiable individuals, or to support population monitoring systems such as disease registries and public health surveillance. Use of health records for other purposes such as teaching, clinical audit, administration and reporting, service management, research and epidemiology, which often require anonymization or aggregation of individual records, are not the focus of ISO 13606-1:2008 but such secondary uses might also find this document useful.

20. ISO 13606-2:2008 [Published]- Electronic health record communication -- Part 2: Archetype interchange specification.

ISO 13606-2:2008 specifies the information architecture required for interoperable communications between systems and services that need or provide EHR data. ISO 13606-2:2008 is not intended to specify the internal architecture or database design of such systems. The subject of the record or record extract to be communicated is an individual person, and the scope of the communication is predominantly with respect to that person's care. Uses of healthcare records for

other purposes such as administration, management, research and epidemiology, which require aggregations of individual people's records, are not the focus of ISO 13606-2:2008 but such secondary uses could also find this document useful. ISO 13606-2:2008 defines an archetype model to be used to represent archetypes when communicated between repositories, and between archetype services. It defines an optional serialized representation, which may be used as an exchange format for communicating individual archetypes. Such communication might, for example, be between archetype libraries or between an archetype service and an EHR persistence or validation service.

21. ISO 13606-3:2008 [Published]- Electronic health record communication -- Part 3: Reference archetypes and term lists.

ISO 13606-3:2009 is for the communication of part or all of the electronic health record (EHR) of a single identified subject of care between EHR systems, or between EHR systems and a centralized EHR data repository. It may also be used for EHR communication between an EHR system or repository and clinical applications or middleware components (such as decision support components) that need to access or provide EHR data, or as the representation of EHR data within a distributed (federated) record system. ISO 13606-3:2009 defines term lists that each specify the set of values that particular attributes of the Reference Model defined in ISO 13606-1 may take. It also defines informative Reference Archetypes that correspond to ENTRY-level compound data structures within the Reference Models of openEHR and HL7 Version 3, to enable those instances to be represented within a consistent structure when communicated using ISO 13606-3:2009.

22. ISO 13606-4:2008 [Published]- Electronic health record communication -- Part 4: Security.

ISO/TS 13606-4:2009 describes a methodology for specifying the privileges necessary to access EHR data. This methodology forms part of the overall EHR communications architecture defined in ISO 13606-1. ISO/TS 13606-4:2009 seeks to address those requirements uniquely pertaining to EHR communications and to represent and communicate EHR-specific information that will inform an access decision. It also refers to general security requirements that apply to EHR communications and points at technical solutions and standards that specify details on services meeting these security needs.

23. ISO 13606-5:2008 [Published]- Electronic health record communication -- Part 5: Interface specification.

ISO 13606-5:2010 specifies the information architecture required for interoperable communications between systems and services that need or provide EHR data. The subject of the record or record extract to be communicated is an individual person, and the scope of the communication is predominantly with respect to that person's care.

ISO 13606-5:2010 defines a set of interfaces to request and provide:

- an EHR_EXTRACT for a given subject of care as defined in ISO 13606-1;

- one or more ARCHETYPE(s) as defined in ISO 13606-2;

- an EHR_AUDIT_LOG_EXTRACT for a given subject of care as defined in ISO/TS 13606-4.

ISO 13606-5:2010 defines the set of interactions for requesting each of these artefacts, and for providing the data to the requesting party or declining the request. An interface to query an EHR or populations of EHRs, for example for clinical audit or research, are beyond its scope, although provision is made for certain selection criteria to be specified when requesting an EHR_EXTRACT which might also serve for population queries. ISO 13606-5:2010 defines the Computational

Viewpoint for each interface, without specifying or restricting particular engineering approaches to implementing these as messages or as service interfaces. ISO 13606-5:2010 effectively defines the payload to be communicated at each interface. It does not specify the particular information that different transport protocols will additionally require, nor the security or authentication procedures that might be agreed between the communicating parties or required by different jurisdictions.

24. ISO/TR 14292:2012 [Published]- Personal health records -- Definition, scope and context.

This Technical Report defines a personal health record (PHR). This definition is intended to help clarify the kinds of records that should be called PHRs, in recognition of the lack of consistency in how this term is presently used. This Technical Report considers the PHR from the perspective of the personal information contained within it and the core services needed to manage this information. A PHR is not a singular entity; the concept encompasses a spectrum of possible information repositories and services that meet different purposes consistent with the definition. This Technical Report therefore also discusses the scope of the PHR in terms of this spectrum as a series of dimensions by which a PHR may be classified and equivalent PHR products compared. It also includes one dimension to classify the kinds of collaborative care PHRs provided by healthcare organisations. This Technical Report also considers the wider context of engagement of individuals in the management of their own health and healthcare, since this engagement is the primary driver for present-day growth of PHR systems and services internationally. This Technical Report includes:

— a definition of a PHR;

— a pragmatic multidimensional classification of PHRs;

— an overview of the possible ways in which the inclusion and engagement of individuals in managing their health and healthcare impacts on the potential roles of the PHR, including scenarios for collaborative care between individuals and healthcare organizations. The many kinds of end-user application that might be implemented and used to deliver PHR system functionality are outside the scope of this Technical Report.

25. ISO/HL7 16527:2016 [Published]- HL7 Personal Health Record System Functional Model, Release 1 (PHRS FM).

ISO/HL7 16527 PHR-S FM:2016 defines a standardised model of the functions that may be present in PHR Systems. It is beyond the scope of the PHR system to control the use (or intended use) of PHR data. On the contrary, it is within the scope of the PHR system to manage the authorisation of an individual (or other application). Those parties are then responsible for using the data for appropriate (or intended) purposes. The system manufacturers specify "intended and permitted use of PHR data" in their Terms of Service and Terms of Use agreements.

The information exchange enabled by the PHR-S supports the retrieval and population of clinical documents and summaries, minimum data sets, and other input/outputs.

26. ISO/TR 17522:2015 [Published]- Provisions for health applications on mobile/smart devices.

ISO/TR 17522:2015 is applicable to the developments of smart health applications available anywhere, anytime and supporting new health businesses based on the smart devices. This Technical Report is to investigate the areas of ongoing developments and analyses of emerging interoperability standards for smart mobile devices.

27. ISO/TS 17975:2015 [Published]- Principles and data requirements for consent in the Collection, Use or Disclosure of personal health information.

ISO/TS 17975:2015 defines the set of frameworks of consent for the Collection, Use and/or Disclosure of personal information by health care practitioners or organizations that are frequently used to obtain agreement to process the personal health information of subjects of care. This is in order to provide an Informational Consent framework which can be specified and used by individual policy domains (e.g. healthcare organisations, regional health authorities, jurisdictions, countries) as an aid to the consistent management of information in the delivery of health care services and the communication of electronic health records across organizational and jurisdictional boundaries. The scope of application of this Technical Specification is limited to Personal Health Information (PHI) as defined in ISO 27799, "information about an identifiable person that relates to the physical or mental health of the individual or to provision of health services to the individual. This information might include:

- information about the registration of the individual for the provision of health services;

- information about payments or eligibility for health care in respect to the individual;

- a number, symbol or particular code assigned to an individual to uniquely identify the individual for health purposes;

- any information about the individual that is collected in the course of the provision of health services to the individual;

- information derived from the testing or examination of a body part or bodily substance;

- identification of a person, e.g. a health professional, as a provider of healthcare to the individual."

Good practice requirements are specified for each framework of Informational Consent. Adherence to these requirements is intended to ensure any subject of care and any parties that process personal health information that their agreement to do so has been properly obtained and correctly specified.

28. ISO 18308:2011 [Published]- Requirements for an electronic health record architecture.

ISO 18308:2011 defines the set of requirements for the architecture of a system that processes, manages and communicates electronic health record (EHR) information: an EHR architecture. The requirements are formulated to ensure that these EHRs are faithful to the needs of healthcare delivery, are clinically valid and reliable, are ethically sound, meet prevailing legal requirements, support good clinical practice and facilitate data analysis for a multitude of purposes. ISO 18308:2011 does not specify the full set of requirements that need to be met by an EHR system for direct patient care or for other use cases, but the requirements defined by ISO 18308:2011 do contribute to the governance of EHR information within such systems.

29. ISO/TR 19231:2014 [Published]- Survey of mHealth projects in low- and middle-income countries (LMIC).

ISO/TR 19231:2014 surveys ongoing national mHealth projects in LMIC, to which some emerging technologies such as zero configuration and proximity computing are applicable, especially when the information and communication technology (ICT) infrastructure is not established in those countries. The scope is constrained to mHealth use cases and technologies for information and communication infrastructures that are useful for LMICs. In addition, the purpose of this Technical Report is to survey not only national mHealth projects in LMICs, but also possible mHealth frameworks that might be used.

30. ISO/TS 20428:2017 [Published]- Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records.

ISO/TS 20428:2017 defines the data elements and their necessary metadata to implement a structured clinical genomic sequencing report and their metadata in electronic health records particularly focusing on the genomic data generated by next generation sequencing technology. These technical specifications:

- defines the composition of a structured clinical sequencing report (see Clause 5),

- defines the required data fields and their metadata for a structured clinical sequencing report (see Clause 6),

- defines the optional data (see Clause 7),

- covers the DNA-level variation from human samples using whole genome sequencing, whole exome sequencing, and targeted sequencing (disease-targeted gene panels) by next generation sequencing technologies. Though whole transcriptome sequencing and other technologies are important to provide better patient care and enable precision medicine, this document only deals with DNA-level changes,

- covers mainly clinical applications and clinical research such as clinical trials and translational research which uses clinical data. However, the necessary steps such as de-identification or consent from patient should be applied. The basic research and other scientific areas are outside the scope of this document,

- does not cover the other biological species, i.e. genomes of viruses and microbes, and

- does not cover the Sanger sequencing methods.

31. ISO/TR 20514:2005 [Published]- Electronic health record -- Definition, scope and context.

ISO/TR 20514:2005 describes a pragmatic classification of electronic health records, provides simple definitions for the main categories of EHR and provides supporting descriptions of the characteristics of electronic health records and record systems.

32. ISO 21549-1:2013 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-1:2013 defines a general structure for the different types of data to be defined in other parts of ISO 21549 using UML notation. ISO 21549 defines data structures held on patient health cards compliant with the physical dimensions of ID-1 cards, as defined by ISO/IEC 7810.

33. ISO 21549-2:2014 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-2:2014 establishes a common framework for the content and the structure of common objects used to construct data held on patient healthcare data cards. It is also applicable to common objects referenced by other data objects. ISO 21549-2:2014 is applicable to situations in which such data is recorded on or transported by patient health cards compliant with the physical dimensions of ID-1 cards defined by ISO/IEC 7810. ISO 21549-2:2014 specifies the basic structure of the data but does not specify or mandate particular data sets for storage on devices.

34. ISO 21549-3:2014 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-3:2014 is applicable to situations in which limited clinical data are recorded on or transported by patient health cards compliant with the physical dimensions of ID-1 cards defined by ISO/IEC 7810. ISO 21549-3:2014 describes and defines the limited clinical data objects used in or referenced by patient health cards using UML, plain text and abstract syntax notation (ASN.1).

ISO 21549-3:2014 specifies the basic structure of the data contained within the data object limited clinical data but does not specify or mandate particular data sets for storage on devices.

35. ISO 21549-4:2014 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-4:2014 is applicable to situations in which clinical data additional to the limited clinical data defined in ISO 21549-3 is recorded on or transported by patient healthcare data cards compliant with the physical dimensions of ID-1 cards defined by ISO/IEC 7810. ISO 21549-4:2014 specifies the basic structure of the data contained within the data object extended clinical data but does not specify or mandate particular data sets for storage on devices.

36. ISO 21549-5:2015 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-5:2015 describes and defines the basic structure of the identification data objects held on healthcare data cards but does not specify particular data sets for storage on devices. The detailed functions and mechanisms of the following services are not within the scope of this part of ISO 21549 (although its structures can accommodate suitable data objects elsewhere specified):

- security functions and related services that are likely to be specified by users for data cards depending on their specific application, e.g. confidentiality protection, data integrity protection and authentication of persons and devices related to these functions;

- access control services;

- the initialization and issuing process (which begins the operating lifetime of an individual data card, and by which the data card is prepared for the data to be subsequently communicated to it according to this part of ISO 21549).

37. ISO 21549-6:2008 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-6:2008 is applicable to situations in which administrative data are recorded on or transported by patient health cards compliant with the physical dimensions of ID-1 cards defined by ISO/IEC 7810. ISO 21549-6:2008 specifies the basic structure of the data contained within the data object administrative data but does not specify or mandate particular data sets for storage on devices. The detailed functions and mechanisms of the following services are not within the scope of this ISO 21549-6:2008, although its structures can accommodate suitable data objects elsewhere specified:

- the encoding of free text data;

- security functions and related services that are likely to be specified by users for data cards depending on their specific application, e.g. confidentiality protection, data integrity protection, and authentication of persons and devices related to these functions;

- access control services that may depend on active use of some data card classes such as microprocessor cards;

- the initialization and issuing process, which begins the operating lifetime of an individual data card, and by which the data card is prepared for the data to be subsequently communicated to it according to this part of ISO 21549.

38. ISO 21549-7:2016 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-7:2016 applies to situations in which such data is recorded on or transported by patient health cards compliant with the physical dimensions of ID-1 cards defined by ISO/IEC 7810. ISO 21549-7:2016 specifies the basic structure of the data contained within the medication data object

but does not specify or mandate particular data sets for storage on devices. The purpose of this document is for cards to provide information to other health professionals and to the patient or its non-professional caregiver. It can also be used to carry a new prescription from the prescriber to the dispenser/pharmacy in the design of its sets. Medication data include the following four components:

- medication notes: additional information related to medication and the safe use of medicines by the patient such as medication history, sensitivities and allergies;

- medication prescriptions: to carry a new prescription from the prescriber to the dispenser/pharmacy;

- medication dispensed: the records of medications dispensed for the patient;

- medication references: pointers to other systems that contain information that makes up medication prescription and the authority to dispense.

ISO 21549-7:2016 describes and defines the Medication data objects used within or referenced by patient-held health data cards using UML, plain text and Abstract Syntax Notation (ASN.1). ISO 21549-7:2016 does not describe nor define the common objects defined within ISO 21549-2, even though they are referenced and utilized within this document.

39. ISO 21549-8:2010 [Published]- Patient health card data -- Part 1: General structure.

ISO 21549-8:2010 defines a way to facilitate access to distributed patient records and/or administrative information using health cards. It defines the structure and elements of "links" typically stored in health cards and representing references to individual patients' records as well as to subcomponents of them. Access control mechanisms, data protection mechanisms, access methods and other security services are outside the scope of ISO 21549-8*:2010.*

40. ISO 25237:2017 [Published]- Pseudonymisation.

ISO 25237:2017 contains principles and requirements for privacy protection using pseudonymisation services for the protection of personal health information. This document is applicable to organizations who wish to undertake pseudonymisation processes for themselves or to organizations who make a claim of trustworthiness for operations engaged in pseudonymisation services. This standard:

- defines one basic concept for pseudonymisation,

-defines one basic methodology for pseudonymisation services including organizational, as well as technical aspects,

- specifies a policy framework and minimal requirements for controlled re-identification (see Clause 7),

- gives an overview of different use cases for pseudonymisation that can be both reversible and irreversible,

- gives a guide to risk assessment for re-identification,

- provides an example of a system that uses de-identification,

- provides informative requirements to an interoperability to pseudonymisation services and

- specifies a policy framework and minimal requirements for trustworthy practices for the operations of a pseudonymisation service.

41. ISO 25720:2009 [Published]- Genomic Sequence Variation Markup Language (GSVML).

ISO 25720:2009 is applicable to the data exchange format that is designed to facilitate the exchange of the genomic sequence variation data around the world, without forcing change of any database schema. From an informatics perspective, GSVML defines the data exchange format based on XML. The scope of ISO 25720:2009 is the data exchange format, but the database schema itself is outside the scope of this International Standard. From a biological point of view, all genetic sequence variations are taken into consideration and are within the scope of this International Standard, while polymorphisms, especially SNP, are the main focus of this International Standard. In other words, the annotations of variation as clinical concerns and -omics concerns are within the scope of ISO 25720:2009. Though SNPs exist in various biological species, the scope of this International Standard covers the human health associated species as human, cell line, and preclinical animals. The other biological species are outside the scope of ISO 25720:2009. The clinical field is within the scope of this International Standard, but the basic research fields and other scientific fields are outside the scope of ISO 25720:2009. Here, clinical research including drug discovery is within the scope of this International Standard. As for supposed application fields, our main focus is in human health including clinical practice, preventive medicine, translational research and clinical researches.

42. ISO/PRF 13606-1 [Under development] - Electronic health record communication -- Part 1: Reference model.
43. ISO/PRF 13606-2 [Under development] - Electronic health record communication -- Part 2: Archetype interchange specification.
44. ISO/PRF 13606-3 [Under development] - Electronic health record communication -- Part 3: Reference archetypes and term lists.
45. ISO/PRF 13606-4 [Under development] - Electronic health record communication -- Part 4: Security.
46. ISO/PRF 13606-5 [Under development] - Electronic health record communication -- Part 5: Interface specification.
47. ISO/AWI TR 20841 [Under development] - Transnational Health Record.
48. ISO/AWI TR 21332 [Under development] - Cloud computing considerations for health information systems security and privacy.
49. ISO/AWI 21393 [Under development] - Omics Markup Language (OML).
50. ISO/DTR 21835 [Under development] - Health-related data which a person generates daily.
51. ISO/AWI 21860 [Under development] - Reference Standards Portfolio for Clinical Imaging (RSP-CI).
52. ISO/AWI TS 22692 [Under development] - Quality control metrics for DNA sequencing.
53. ISO/AWI TS 22693 [Under development] - Structured clinical gene fusion report in electronic health records.
54. ISO/AWI 25720 [Under development] - Genomic Sequence Variation Markup Language (GSVML).
55. ISO/TR 14639-1:2012 [Published]- Capacity-based eHealth architecture roadmap -- Part 1: Overview of national eHealth initiatives.

ISO/TR 14639-1:2012 aims to identify the business requirements of an eHealth architecture as well as providing a generic and comprehensive context description to inform architectural structuring of Health Information Systems (HIS). ISO/TR 14639-1:2012 reviews international experiences in the construction of national eHealth architectures and introduces a methodology for strategic development of HIS.

56. ISO/TR 14639-2:2014 [Published]- Capacity-based eHealth architecture roadmap -- Part 2: Architectural components and maturity model.

ISO/TR 14639:2014 provides a guide to best practice business requirements and principles for countries and their subordinate health authorities planning and implementing the use of information and communications technology (ICT) to support the delivery and development of healthcare. A business reference architecture is described in terms of components and capabilities that health authorities may use as a framework for building their own eHealth architectures and also for measuring the maturity of their health systems' use of ICT to support the delivery and development of healthcare. This Technical Report also proposes a maturity model and methodology that organizations may consider in developing and evolving their eHealth capacities in specified areas of operational capability from low to medium to high levels. The proposed business reference architecture identifies components and capabilities needed to support various health service activities along with the governance, infrastructure, and ICT infrastructure that is necessary for the effective and efficient use of information in the delivery and development of health services.

57. ISO 22857:2013 [Published]- Guidelines on data protection to facilitate trans-border flows of personal health data.

ISO 22857:2013 provides guidance on data protection requirements to facilitate the transfer of personal health data across national or jurisdictional borders. It is normative only in respect of international or trans-jurisdictional exchange of personal health data. However, it can be informative with respect to the protection of health information within national/jurisdictional boundaries and provide assistance to national or jurisdictional bodies involved in the development and implementation of data protection principles. It covers both the data protection principles that apply to international or trans-jurisdictional transfers and the security policy which an organization adopts to ensure compliance with those principles. It aims to facilitate international and trans-jurisdictional health-related applications involving the transfer of personal health data. It seeks to provide the means by which health data relating to data subjects, such as patients, will be adequately protected when sent to, and processed in, another country/jurisdiction.

b) Under the umbrella of IEEE:

1. 1926.1 [Under development] - Standard for a Functional Architecture of Distributed Energy Efficient Big Data Processing.

This standard improves the energy efficiency of data networks involved in the processing and transmission of big data. Dealing with the large data volumes generated by big data applications requires a mechanism to handle the trade-offs between transmission and processing from an energy consumption viewpoint.

2. 1752 [Under development] - Standard for Mobile Health Data.

The purpose is to provide standard semantics to enable meaningful description, exchange, sharing, and use of mobile health data. Data and associated metadata will be sufficiently clear and complete to support analysis for a set of consumer health, biomedical research, and clinical care

needs. This standard will leverage data and nomenclature standards such as the IEEE 11073 family of standards for personal health devices as references.

3. 1912 [Under development] - Standard for Privacy and Security Architecture for Consumer Wireless Devices.

This standard describes a common communication architecture for diverse wireless communication devices such as, but not limited to, devices equipped with near field communication (NFC), home area network (HAN), wireless area network (WAN) wireless personal area network (WPAN) technologies or radio frequency identification technology (RFID) considering proximity; and specifies approaches for end user security through device discovery/recognition, simplification of user authentication, tracking items/people under user control/responsibility, and supports alerting; while supporting privacy through user controlled sharing of information independent of the underlying wireless networking technology used by the devices.

c) Under the umbrella of CDSIC[91]:

1. Standard for Exchange of Nonclinical Data (SEND)

SEND is an implementation of the SDTM standard for nonclinical studies. SEND specifies a way to collect and present nonclinical data in a consistent format.

SENDING 3.1 is the CDISC Standard for Exchange of Nonclinical Data Implementation Guide (SENDIG) for nonclinical studies in the context of drug development, which has been prepared by the Standard for Exchange of Nonclinical Data (SEND) Team of the Clinical Data Interchange Standards Consortium (CDISC). The SENDIG is intended to guide the organization, structure, and format of standard nonclinical tabulation datasets for interchange between organizations such as sponsors and CROs, and for submission to regulatory authorities.

The SENDIG is based upon and should be used in close concert with Version 1.5 of the CDISC Study Data Tabulation Model (SDTM).

2. Protocol Representation Model (PRM)

PRM provides a standard for planning and designing a research protocol with focus on study characteristics such as study design, eligibility criteria, and requirements from the ClinicalTrials.gov, World Health Organization (WHO) registries, and EudraCT registries. PRM assists in automating case report forms (CRF) creation and EHR configuration to support clinical research and data sharing.

PRM v1.0 was developed to support: a) protocol document generation; b) research study (clinical trial) registration and tracking; c) regulatory oversight and review; and d) single-sourced, downstream electronic consumption of protocol content, allowing users to create and quality control content once, and reuse for trial registries, protocol and case study report templates, SDTM study design and more.

3. Study/Trial Design Model-XML

Study/Trial Design Model in XML (SDM-XML) is an extension of ODM-XML and allows organisations to provide rigorous, machine-readable, interchangeable descriptions of the designs of their clinical studies, including treatment plans, eligibility and times and events. SDM-XML defines three key

[91] CDISC standards are mandated by FDA and PDMA and recommended by EMA

sub-modules – Structure, Workflow, and Timing – permitting various levels of detail in any representation of a clinical study's design.

4. Clinical Data Acquisition Standards Harmonization (CDASH)

CDASH v2.0 establishes a standard way to collect data in a similar way across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into the Study Data Tabulation Model (SDTM), delivering more transparency to regulators and others who conduct data review.

5. Study Data Tabulation Model (SDTM)

SDTM v1.6 provides a standard for organising and formatting data to streamline processes in collection, management, analysis and reporting. Implementing SDTM supports data aggregation and warehousing; fosters mining and reuse; facilitates sharing; helps perform due diligence and other important data review activities; and improves the regulatory review and approval process. SDTM is also used in non-clinical data (SEND), medical devices and pharmacogenomics/genetics studies. SDTM v1.6 support the SDTM Implementation Guide(SDTMIG) version 1.1 (SENDIG DART v1.1)

6. Analysis Data Model (ADaM)

ADaM defines dataset and metadata standards that support:

- efficient generation, replication, and review of clinical trial statistical analyses, and,
- traceability among analysis results, analysis data, and data represented in the Study Data Tabulation Model (SDTM).

ADaM implementation guide (ADaMIG) v 1.1 updates Version 1.0 with clarifications, corrections, new variables, additional examples, and references to current documents. It is intended to guide the organisation, structure, and format of analysis datasets and related metadata. ADaMIG v 1.1 specifies ADaM standard dataset structures and variables, including naming conventions, and presents standard solutions to implementation issues, illustrated with examples. The ADaMIG must be used in close concert with the ADaM document v2.1.

7. Operational Data Model (ODM)-XML

ODM-XML v1.3.2 is a vendor-neutral, platform-independent format for exchanging and archiving clinical and translational research data, along with their associated metadata, administrative data, reference data, and audit information. ODM-XML facilitates the regulatory-compliant acquisition, archival and exchange of metadata and data. It has become the language of choice for representing case report form content in many electronic data capture (EDC) tools.

8. Dataset-XML

Dataset-XML (v1.0) supports exchanging tabular data in clinical research applications using ODM-based XML technologies, enabling the communication of study datasets for regulatory submissions.

CDISC developed Dataset-XML v1.0 as a drop-in replacement for SAS V5 XPORT to enable testing using existing processes. Dataset-XML is a truly non-proprietary, global standard, removing many SAS V5 Transport file restrictions (the current file format required by the FDA and PMDA), such as 8-character variable names and 200-character text fields. Dataset-XML and Define-XML are complementary standards; Define-XML metadata describes the Dataset-XML dataset content. Dataset-XML can represent any tabular dataset including SDTM, ADaM, SEND, or non-standard legacy datasets. It requires an accompanying Define-XML file to provide the machine-readable

metadata that describes the contents of the datasets. Dataset-XML supports all language encodings supported by XML. Now that Dataset-XML v1.0 has been shown to work as a SAS V5 XPORT replacement, the CDISC XML Technologies Team will add additional features in the next versions, including improved relationships and traceability.

9. Clinical Trial Registry XML (CTR-XML)

CTR-XML lets technology vendors implement tools that support a "write once, use many times" solution based on a single XML file that holds the information needed to generate submissions for multiple clinical trials for clinical trial registry submissions primarily to the World Health Organization (WHO), European Medicines Agency (EMA) EudraCT Registry and United States ClinicalTrials.gov.

10. Resource Description Framework (RDF)

CDISC Standards in RDF provides a representation of the CDISC Foundational standards[92] in a model based on the Resource Description Framework (RDF). RDF provides executable, machine-readable CDISC standards from CDISC SHARE. This file format is a "linked data" view of the standards as an ontology.

Version 1.0 of the CDISC Standards in RDF, prepared by the PhUSE CS Semantic Technology Working Group, consists of two documents:

- CDISC Standards in RDF Reference Guide v1 Final - provides a reference to the representation of the existing foundational CDISC standards in a model based on the Resource Description Framework (RDF).
- CDISC Standards in RDF User Guide v1 Final – describes how to access and use the RDF files and provides background on their creation.

11. Therapeutic Areas

Therapeutic Area (TA) Standards extend the Foundational Standards to represent data that pertains to specific disease areas. TA Standards include disease-specific metadata, examples and guidance on implementing CDISC standards for a variety of uses, including global regulatory submissions.

12. Controlled Terminology

CDISC Controlled Terminology is the set of CDISC-developed or CDISC-adopted standard expressions (values) used with data items within CDISC-defined datasets. CDISC, in collaboration with the National Cancer Institute's Enterprise Vocabulary Services (EVS), supports the controlled terminology needs of CDISC Foundational and Therapeutic Area Standards.

d) Under the umbrella of HL7:

1. ISO/HL7 10781:2015

Health Informatics -- HL7 Electronic Health Records-System Functional Model, Release 2 (EHR FM) provides a reference list of functions that may be present in an Electronic Health Record System (EHR-S). The function list is described from a user perspective with the intent to enable consistent expression of system functionality. This EHR-S Functional Model, through the creation of Functional Profiles for care settings and realms, enables a standardized description and common

[92] https://www.cdisc.org/standards/foundational

understanding of functions sought or available in a given setting (e.g. intensive care, cardiology, office practice in one country or primary care in another country).

2. ISO/HL7 16527:2016

Health informatics -- HL7 Personal Health Record System Functional Model, Release 1 (PHRS FM).

ISO/HL7 16527 PHR-S FM:2016 defines a standardized model of the functions that may be present in PHR Systems.

Within the scope of the PHR system is to manage the authorisation of an individual (or other application). Those parties are then responsible for using the data for appropriate (or intended) purposes. The system manufacturers specify "intended and permitted use of PHR data" in their Terms of Service and Terms of Use agreements. The information exchange enabled by the PHR-S supports the retrieval and population of clinical documents and summaries, minimum data sets, and other input/outputs.

3. ISO/HL7 21731:2014

Health informatics -- HL7 version 3 -- Reference information model -- Release 4

4. ISO/HL7 27932:2009

Data Exchange Standards -- HL7 Clinical Document Architecture, Release 2

ISO 27932:2009 covers the standardization of clinical documents for exchange.

5. ISO/HL7 27953-1:2011

Health informatics -- Individual case safety reports (ICSRs) in pharmacovigilance -- Part 1: Framework for adverse event reporting

ISO 27953-1:2011 seeks to establish an international framework for data exchange and information sharing by providing a common messaging format for transmission of ICSRs for adverse drug reactions (ADR), adverse events (AE), product problems and consumer complaints that can occur upon the administration or use of one or more products. The messaging format is based upon the HL7 Reference Information Model (RIM) and can be extended or constrained to accommodate a variety of reporting use cases. ISO 27953-1:2011 will be harmonized over time with other HL7 public health and patient safety reporting standards to help ensure that messaging constructs and vocabulary are harmonized in the HL7 Public Health and Regulatory Reporting domains. The data elements used in ISO 27953-1:2011 were identified as consistent across many of the use cases and can be applied to a variety of reporting scenarios. Specific reporting requirements within organizations or regions might vary.

6. ISO/HL7 27953-2:2011

Health informatics -- Individual case safety reports (ICSRs) in pharmacovigilance -- Part 2: Human pharmaceutical reporting requirements for ICSR ISO 27593-2:2011 seeks to create a standardized framework for international regulatory reporting and information sharing by providing a common set of data elements and a messaging format for transmission of ICSRs for adverse drug reactions (ADR), adverse events (AE), infections, and incidents that can occur upon the administration of one or more human pharmaceutical products to a patient, regardless of source and destination.

6. Annex B

6.1. Types of machine learning algorithms

The three major categories of machine learning algorithms are defined as follows:

- *Supervised learning*: it involves building a statistical model when both the input and output are available. In supervised learning for image processing, for example, pictures of vehicles are provided (the input) as well as their categories such as cars and trucks (the output or label). The goal is to predict the output variables or to infer how the output is affected by the input variable as accurately as possible.

  It is called supervised because of the presence of the outcome variable to guide the learning process that can be thought of as a teacher supervising the learning process. The correct answer is known and the algorithm iteratively makes predictions on the input data available; learning to stop when it achieves an acceptable level of performance.

  Supervised learning problems can be further grouped into regression and classification problems:

  - Classification: a classification problem is when the output variable is a category, such as "disease" and "no disease".
  - *Regression*: a regression problem is when the output variable is a real value, such as "value" or "count".

  Algorithms in this category include (not limited to):

  - Linear and logistic regression.
  - Ensemble techniques (including random forests).
  - Deep or shallow neural nets.
  - Recurrent neural networks (including LSTM and GRU).
  - Decision trees.
  - Support vector machines and regression.
- *Unsupervised learning*: it involves a set of algorithms where only input data is available and no corresponding output. The goal is to *infer* a conclusion or hidden structure, to describe how the data are organised or clustered.

  These algorithms are called unsupervised learning because unlike supervised learning there are no correct answers and there is no teacher. Algorithms are left to their own devices to discover and present the patterns in the data, roughly meaning letting them figure things out by themselves.

  These problems are much more common in real world data situations.

  Unsupervised learning problems can be further grouped into three types:

  - Association: the aim is to discover what variables tend to occur together, such as people that take X also tend to experience Z. As an example: if a customer purchases bread, he is 80% likely to also purchase eggs.

- Clustering: the aim is to discover the inherent groupings in the data, such that objects or observations within the same cluster are more similar to each other than to the objects from another cluster.
- Dimensionality reduction (embedding): the aim is to reduce the number of variables of a dataset while ensuring that important information is still conveyed. This can be done using feature selection to selects a subset of the most important original variables, or feature extraction methods to perform data transformation from a high-dimensional space to a low-dimensional space.

Algorithms in this category include (not limited to):

- A priori algorithm for association rule learning problems.
- Hierarchical or K-means clustering.
- Principal and independent component analysis.
- Feature selection algorithms.
- Word2Vec and FastText.
- Anomaly detection.
- Auto-encoders and restricted Boltzmann machines (RBM).

- *Reinforcement learning*: a set of trial-and-error-based algorithms that observes the environment and discovers the best next action by trying to maximise rewards it receives. Essentially machines can learn from and remember the best action by correcting itself over time (see Figure 21). Differently from supervised learning, it does not assume labelled input/output pairs for training. Instead it makes use of available knowledge (exploitation) to explore unknown portions of the search space (exploration). They are usually used in recommendation systems such as several online recommendation systems that suggest users what they might like to buy or also in marketing.

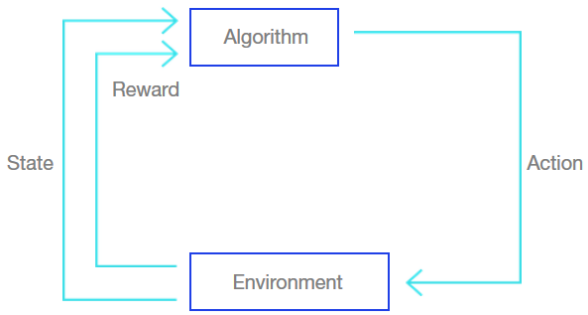


Figure 21: Reinforcement learning.

* Source: McKinsey and Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction, Second edition

Reinforcement learning does not belong with supervised learning, as the data is unlabelled; on the other hand, it is not unsupervised learning either as the algorithm receives some feedback environment.

Google DeepMind has used reinforcement learning to develop systems that can play games, including video games and board games such as Go, better than human champions. Reinforcement

learning requires a lot of data for training and, with games, simulated data is readily available. Other famous applications are stock trading bots where the software observes stock market and the objective is to choose buy/ sell actions in order to maximize the wealth. They are typically used in robotics – where a robot can learn to avoid collisions by receiving negative feedback after bumping into obstacles, and in video games – where trial and error reveals specific movements that can shoot up a player's rewards[93].

Algorithms in this category include:

- Temporal difference learning.
- Q-learning.
- Learning automata.

Sometimes semi-supervised learning algorithms are also mentioned as another type of algorithm. As the name suggests, these sit in between both supervised and unsupervised learning and are characterised by a large amount of input data with only some of them with the corresponding outcome available. Many real-world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts, whereas unlabelled data is cheap and easy to collect. A good example is a photo archive where only some of the images are labelled and the majority are not.

Another important classification of machine learning approach is between parametric and non-parametric methods:

- *Parametric* methods assume a functional form, or shape, of the data. For example, one very simple assumption is that the functional form is linear as in:

  $f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

  This is an example of a linear model.

  After a model or functional form has been selected, a procedure is then needed to estimate the parameter $\beta 0, \beta 1, \ldots, \beta p$ such that the predicted value are as close as possible to the training data.

  Parametric models have several advantages; they are often easy to fit, because only a small number of coefficients need to be estimated. But they do have a disadvantage: by construction, they make strong assumptions about the form of f(X). If the specified functional form is far from the truth, and prediction accuracy is the goal, then the parametric method will perform poorly. For instance, if a linear relationship is assumed between the predictors and the outcome but the true relationship is far from linear, then the resulting model will provide a poor fit to the data.

- *Non-parametric* methods do not assume a functional form of the data, and thereby provide an alternative and more flexible approach having the potential to accurately fit a wider range of possible shapes. But non-parametric approaches do suffer from a major disadvantage: since they do not have a functional form of the data, they need to estimate a larger number of parameters, and they require a very large number of observations (far more than is typically needed for a parametric approach).

[93] Artificial Intelligence: implications for business strategy. MIT Management Executive Education online course 2018

6.2. Deep learning

Artificial Neural Networks (ANN) are inspired by the behaviour of human brain. They are built from artificial neurons (nodes) which are simplified versions of the biological neurons in the mammalian brain. Artificial neurons are stimulated by input signals and they pass on some of this information to other nodes. Similarly, biological neurons the artificial neurons can be trained to pass forward only important signals. The goal of training the network is that only relevant input signals are transformed and passed on to the following neurons.

In the neural network nodes are connected to other nodes. In a deep network the nodes are organised in several layers. The connection weights between the nodes are the primary mean of storing information in NN and during the training these weights are updated as the NN learns new information (Patterson & Gibson, 2017). From the mathematical point of view ANNs are graphs. They describe mathematical processes using hierarchical structure, which allows deep ANNs to represent very complicated mapping from inputs to outputs. In many cases, real world data is also structured hierarchical way and deep ANNs automatically take advantage of this fact.

Figure 22 presents a single artificial neuron. Each neuron takes the input from incoming connections (arrows = edges of the mathematical graph) and passes that on to its' own activation function (f in Figure below). The activation function is used to produce neurons output from the inputs.

The net input from each connection to the neuron is formed by adjusting the input by a weight parameter (wi) on that connection. The weight parameter either amplifies or minimises the original input signal. The goal of training the model is to find such weights that amplify the signal and dampen the noise. The weighted sum of inputs is further adjusted by adding so called bias term to it. The biases ensure that at least some nodes per layer are activated regardless of signal strength and thus they help in the event of low signal.

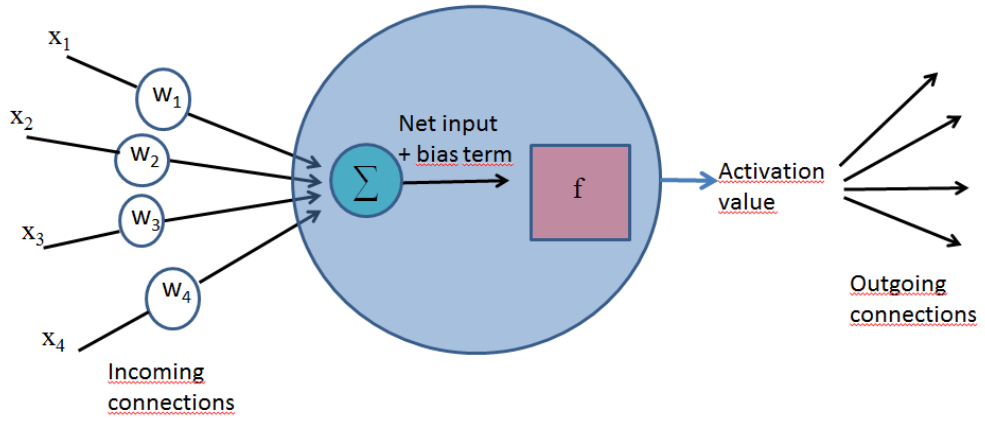
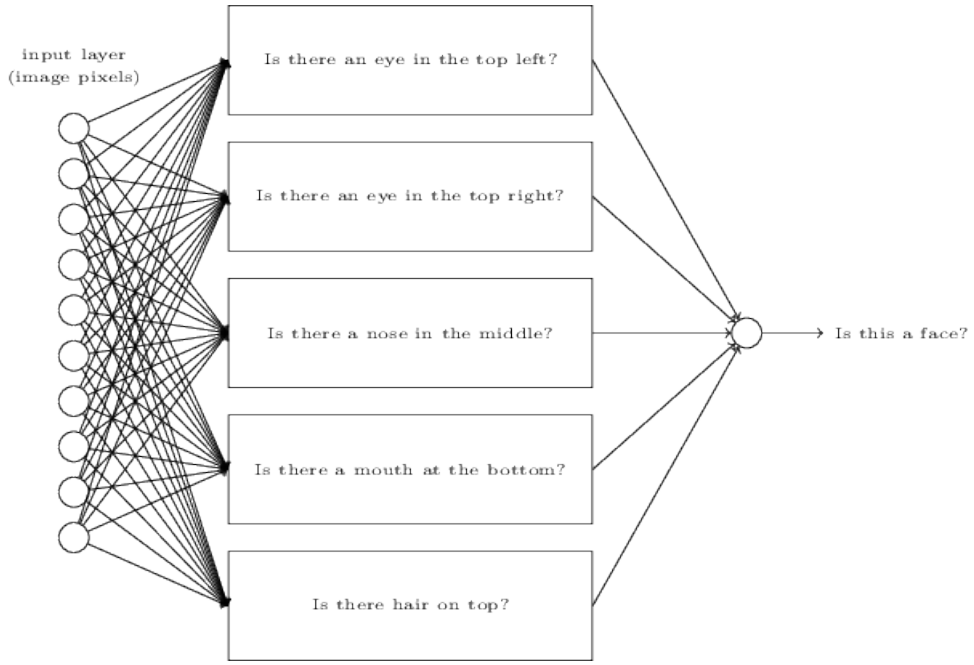


Figure 22: Artificial neuron

*Source: Patterson & Gibson 2017



During the training stage, the weight and bias parameters are optimised in such a way that the model minimises a loss function (Patterson & Gibson, 2017). The loss-function is based on some measure of the difference between the network outputs and the desired outputs.

One important factor contributing to the increased computing power is the evolution of parallel computing. As ANNs can be viewed as graphs, it is easily to break up the graph into parts, where the training computations can be performed simultaneously (parallel) rather than sequentially. Thus, the training of ANN is a task, which parallelises well (not all computing tasks are such that parallel computing increases the speed dramatically).

An important step forward in parallel computing has been the utilization of Graphical Processing Units (GPUs). GPU consists of hundreds or even thousands of processor cores and they are very powerful when harnessed to the tasks of parallel computing. The development of GPUs originates from gaming industry. GPUs have provided access to advantages of accelerated computing with low cost compared to traditional supercomputers. GPUs have thus made deep ANN training available practically to anyone.

Automatic feature extraction is one of the most important advantages of deep learning. When thinking about building a classifier model, researchers have traditionally spent much of their time figuring out and choosing features, which are distinctive between different classes. Building this kind of handcrafted model typically requires field specific expertise. Deep learning model are different in this respect as they select the most important distinctive features automatically from the input data given to them. Therefore, deep learning models are considered quite easy to apply as no field specific knowledge is required. The other side of the coin is that even though the classifier based on deep learning may perform well, one cannot say which features of the data the model actually uses. In other words, NN models are black boxes as they are too complicated for human to understand. The number of parameters may be hundreds or thousands and the causal logic behind the model conclusions cannot be seen.

Medicinal industry is rapidly becoming very data intensive and frequently the data is complex. Moreover, the data is often not well understood, and hence deep learning methods have promising potential in this field (Ching et.al 2018). Potential applications exist in the areas of translational bioinformatics, medical imaging, pervasive sensing, medical informatics, and public health. Figure 23 below shows that the number of scientific publications is growing rapidly in 2010s and that the imaging

analysis applications have thus far been the most common subarea of deep learning in health informatics (Radi et.al 2017).

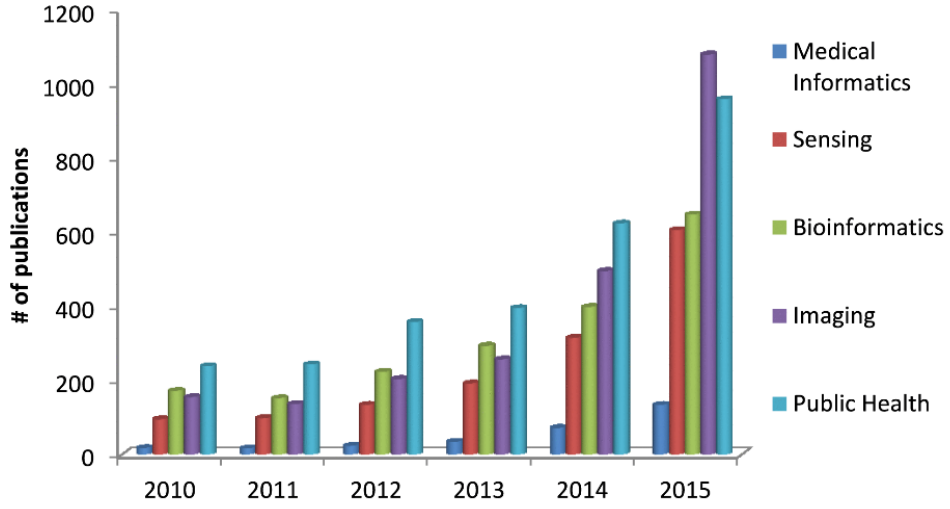


Figure 23: Distribution of published papers that use deep learning in subareas of health informatics

Source: Radi et.al. 2017

Deep learning methods are highly scalable and therefore suitable for analysis of large datasets. One can also say that to yield useful results, deep learning methods require a large set of training data and without such data they are not applicable. Therefore, first applications of deep learning have been directed to the fields, where suitable data are readily available (picking low-hanging fruits). For example, the amount of -omics data is growing rapidly, and this has become a field, where deep learning is frequently applied. Furthermore, as most of the current deep learning applications belong in the category of supervised learning, the data for model training needs to be labelled. In some cases, the lack of enough labelled data prevents researchers tackling the most interesting questions by deep learning methods.

The flexibility of ANN model is partly due to fact that there are many hyperparameters to adjust in these models. Hyperparameters are parameters whose value is set before the learning process begins. They are used to control model structure (number of neurons/ layers), the learning process itself, initialisation of the model and the properties of the neuron activation function etc. Traditionally tuning of the hyperparameters is done by trial and error, but also modern complex algorithms help this tuning have become available[94].

6.2.1. Multilayer Feed-Forward Network

The multilayer feed-forward network is the earliest model utilising deep learning. The network is composed of one input layer, multiple hidden layers and finally the output layer (see Figure 24). As the data moves forward in the network it is influenced by connection weights. Moreover, the neurons filter the data they receive, and aggregate, convert, and transmit only certain information to the next neuron (Patterson & Gibson, 2017).

[94] Sometimes a third dataset, besides training and test sets, is formed. This is called **validation set** and it used for tuning the hyperparameters

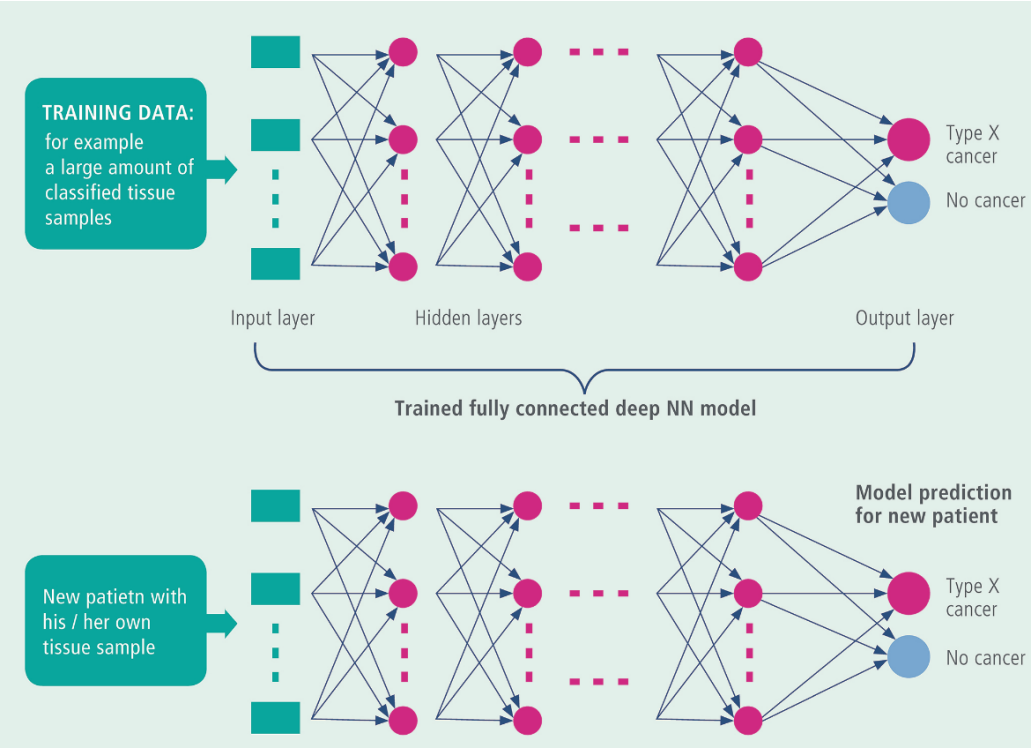


Figure 24: Simplified schema of fully connected deep neural network

6.2.2. Convolutional Neural Networks

Deep Convolutional Neural Networks (CNNs) have dominated image recognition challenges in 2010s and their success is a significant contributor to the creation of the deep learning hype. Deep CNNs are also utilised extensively in the health care sector for image analysis or more generally for object recognition and detection.

CNN architectures have several variations, but in general CNN are similar to the multilayer feed-forward networks described above except that they are not a fully connected network: some neurons are only locally connected.

In a fully connected network, each neuron is connected to a neighbourhood of neurons in the previous layer. In CNN the neurons are only connected to a subset of the input and this allow the number of parameters in the model to reduce significantly. Moreover, CNNs usually further constrain the number of parameters by using a parameter sharing scheme (not all parameters are fully adjustable, but are adjusted together), which also helps to reduce the training time (Patterson & Gibson, 2017).

Having neurons connected only to a subset of the input also helps to control overfitting and provides the network "funnel like" shape by shrinking the size. This can represent an advantage in case the data have some structure and the subset of connection acts as a feature extractor. The local connectivity of early convolutional layers allows the network to learn low-level features of the input data. Subsequent layers then assemble these features into higher level features.

The CNNs tolerate location invariance better than most other methods in image recognition tasks. A particular object can appear in various locations in the image and its rotation and scale may vary. These facts pose a difficult challenge for most automated image analysis tools.

6.2.3. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are particularly well suited for modelling time series data or more generally sequences, including text. In contrast to previous feed-forward networks, they also have connections between the time steps. The simplest example is recurrent neuron, which in addition to other inputs, receives its own output from previous time step as an input. Figure 25 presents this kind of network that has been "unrolled" in time dimension. One can observe the connection from hidden layer at time t-1 to hidden layer at time t. Regular feed-forward networks implicitly assume time independence of input and/or output. RNNs in contrast allow input and/or output vectors whose values are time dependent (Patterson & Gibson, 2017).

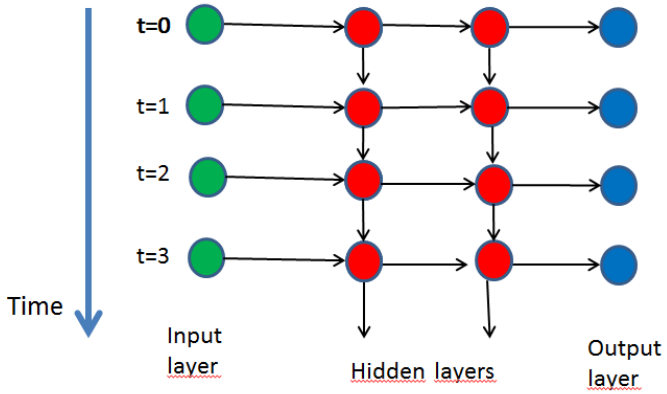


Figure 25: Recurrent Neural Network unrolled along the time axis

Source: Patterson & Gibson, 2017

The most commonly used RNN type is so called Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997). Natural language processing is popular application for RNNs (such as speech transcription to text and machine translation). Also, the analysis of sequences such as stock prices or trajectories of objects (within autonomous driving) are suitable for RNNs.

6.2.4. Autoencoders

Autoencoders are unsupervised learning algorithms and they use unlabelled data. An autoencoder is a type of neural network that contains a hidden layer that has smaller dimension than either input or output layers. In a sense the data gets compressed as it is passed through this layer (see Figure 26). This phase is called encoding and the decoding phase is happening when the data is reconstructed again in output layer. The purpose of autoencoder training is to learn parameters that make the output as close to the input as possible.

As the data is compressed in encoding phase, autoencoders can be used in dimensionality reduction of the data or for de-noising the data. In that case, the purpose is to learn how to produce uncontaminated output from contaminated input data.

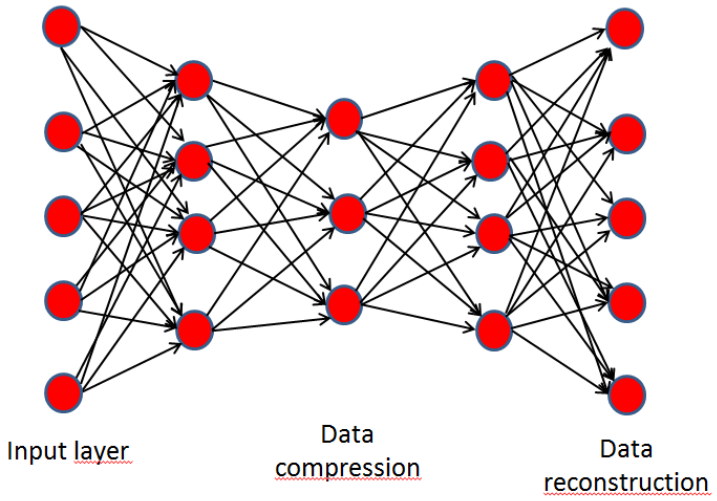


Figure 26: Autoencoder network architecture

Source: Patterson & Gibson, 2017

6.2.5. Conclusions on deep learning

Deep learning is the core of current AI hype. Deep learning has a lot of potential in the field of pharma and medicine industry. However, some of the enthusiasm may be misplaced at this time. They have just recently become available for practical application. So, there are many application domains waiting to be found.

The nature of most deep learning applications has been exploratory, or hypothesis generating, analysis. This means that one still has to confirm the findings using traditional methods such as randomized clinical trials. Deep ANNs are black box methods and understanding how users should interpret these models is not clear. In many cases understanding the causal logic behind the results and generating testable hypotheses is more interesting than the results by themselves. Large effort has been dedicated to research of making ANNs more interpretable, but thus far a breakthrough in this area has not been made.

Deep learning methods and tools to implement them are in the phase of very rapid progress. Scientist and professionals have to struggle to keep up in this development. Consequently, no uniformly accepted standards on methods or tools exist.

Above several different ANN architectures were described and there exist still plenty more. An important point in applications is to match the model's architecture to the input data. Network's architecture is not restricted to single type, but combination or hybrid models of different types can be developed.

6.3. Time-series

A time series is a sequence of observations on a variable that is measured over successive points in time. These points in time may be equal (e.g. daily) or non-equal. A time series can be considered a particular realisation of a stochastic process, i.e. a collection of random variables, $\{x_t\}$, indexed by time t.

Characteristically, time series are often structured so that observations are nested within larger units of time, such that, for instance, observations collected on a daily basis are nested within months or years.

Time-series models have not been extensively used in drug regulation, when compared to other types of epidemiological and analytical approaches. However, this is changing, with an increased number of scientific publications exploring the use of time-series in real-world data in the context of drug regulation.

In fact, time-series may provide valuable information that can be used in the regulatory context, particularly when using observational data. For instance, an unexpected rise in the frequency of an adverse drug reaction may indicate the presence of a quality defect. Similarly, the trend-change in a time series after the implementation of a risk minimisation measure may give an indication of the impact of the measure.

Several areas of scientific knowledge make ample use of time-series analyses, such as geography, finance, economics and epidemiological surveillance among others. This has resulted in a wide array of off-the-shelf methods, many of which freely available as open-source software.

It is, however, important to consider the specificities of time-series and the characteristics of the variety of models available for running time-series analyses and understand their use and limitations, prior to implementing them as a regulatory tool.

This sub-section reflects on basic concepts and potential uses of time-series analysis in drug regulation, particularly as they apply to non-interventional data. An in-depth discussion on the underlying statistics is beyond the scope.

6.3.1. Purpose for conducting a time-series analysis

There are roughly four objectives to conducting time series analyses: describing the time series, identifying relationships, forecasting and analysing interventions.

Describing the characteristics of the time series involves identifying and characterising patterns in the data, such as trend and autocorrelations. Identifying relationships is performed by assessing the measure of similarity of the change in one variable compared to the change in another, i.e. cross-correlation. Forecasting is the process of making predictions based on the observed series. Finally, time series analyses can be performed to understand the impact of an intervention, such as a public health programme.

6.3.2. Stationarity

A stochastic process is said to be stationary if its statistical characteristics do not change over time. Stationarity can be considered a horizontal pattern, where the data fluctuates around a constant mean however simply observing a horizontal pattern is not sufficient evidence to conclude that the time series is stationary.

Stationarity is important insomuch as most statistical methods to analyse time series are based on the assumption that time series are either stationary or can be rendered approximately stationary through some transformation. A stationary time series is relatively easy to characterise and predict, as its statistical properties will remain unchanged over time.

6.3.3. Time-series variation

Seasonality and cycles

Seasonal and cyclical variation are characterised by repeating fluctuations over the time span of the observations. The distinction is subtle. Seasonal variations are variations in the observations of a variable at regular and predictable intervals that recur within a calendar year, such as weekly, monthly or quarterly. Cyclical variations are regular oscillations in the observations of a variable that occur periodically but not within a calendar period, such as a 3-year cycle.

Noticeably, the magnitude of the changes may vary across periods, but the duration and direction are the same on each occasion.

Trend

A trend is loosely a long-term change in the mean level. The challenge with this definition is deciding what long-term is, as the length of observations may influence the trend. For instance, a yearly increase in observations may actually be an overall decrease in the observations over ten years.

The direction and slope of a trend can remain constant or change throughout the series. Modelling an observed trend over time using regression is appropriate when the trend is deterministic, i.e. due to the deterministic effects of some causal agent. Because of this causal structure, a deterministic trend is generally stable across time. Conversely, a time series can also exhibit stochastic trends, which can arise simply from the random movement of a variable over time.

One of the challenges is identifying the period in which a causal agent exerts its effect, and thus where the trend is deterministic. Consider for instance the assessment of the impact, i.e. the deterministic effect, of two regulatory interventions: for the withdrawal of a product with a recall, the effect is near immediate and measured within shorter units of time (e.g. weeks), whereas the deterministic effect of the introduction of a new contra-indication may only be observable after months or years.

In the second example, the trend between the time of regulatory action and the effect of the measure could be a stochastic trend or still influenced by the antecedent deterministic effect, i.e. the risk.

6.3.4. Autocorrelation

There may be a degree of similarity between a given time series and a lagged version of itself, such that observed values of a variable X at a given time $_i$, X_i, may be correlated with X_{i-k}, a lag of unit k. Where this occurs, the time series may be said to present autocorrelation, or serial correlation.

One way of assessing autocorrelation is by plotting the autocorrelation function (ACF) of a time series. Autocorrelation plots indicate the correlation of a variable and a sequence of lags of itself and can range from - 1 to 1. An example of autocorrelated data is the monthly antidiabetic drug prescription data (pharmaceutical products of the ATC code A10) from the Australian Health Insurance, from 1992 to 2007. As the data is grouped by months, the ACF plot x-axys shows one calendar cycle as a year, thus 1 refers the same month of the year in subsequent years. When lag k is 0 the correlation is 1 and naturally decreases from there but still showing a seasonal pattern.

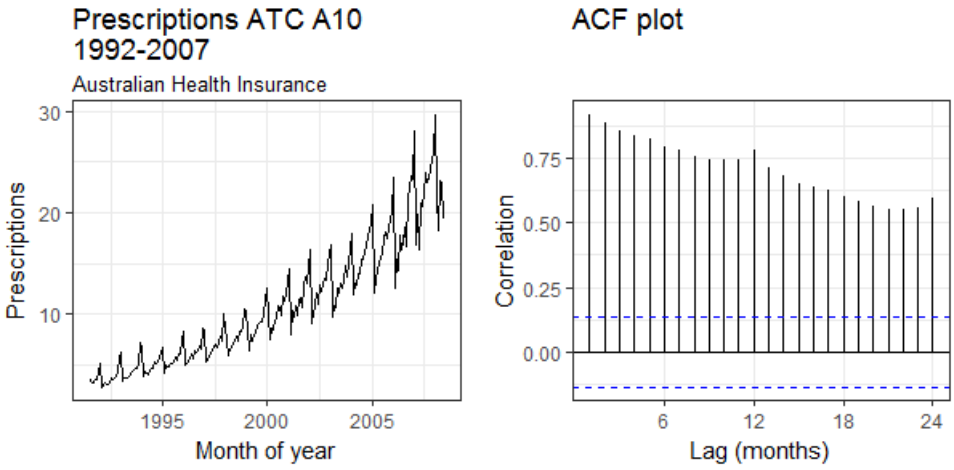


Figure 27: Example of autocorrelated data

Autocorrelation is an important aspect of time series insomuch as its presence may invalidate statistical principles of certain analyses. In fact, normal regression methods assume independence of errors, autocorrelation violates this.

6.3.5. Decomposition

At its simplest, a time-series can be considered as comprising four components[95], a trend (T), a cycle (C), a seasonal component (S) and a remainder or random noise (N). Two of the more important decomposition methods are the additive decomposition and the multiplicative decomposition.

The additive decomposition is expressed as the arithmetic sum of each of the components, such that $y_t = T_t + C_t + S_t + N_t$, whereas the multiplicative decomposition model is expressed as the product of the components, $y_t = T_t.C_t.S_t.N_t$ or $\log y_t = \log T_t + \log C_t + \log S_t + \log N_t$.

The additive model is the most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend and cycle does not vary with the level of the time series. While this difference may not be obvious in visual interpretations of the time-series plot, the air passenger plot above clearly suggests a multiplicative model.

Classical decomposition methods, both for additive and multiplicative decomposition are widely used but are limited in that they assume a seasonal component that is constant from year to year. More recent methods of decomposition include X-12-ARIMA and STL (seasonal and trend decomposition using Loess[96]).

Decomposition is helpful to obtain seasonally adjusted data, particularly where the analytic method to be applied does not take into account an autocorrelation structure.

6.3.6. Transformations

A generic approach to data transformations is described in a separate chapter. There are, however, certain data transformations that are specific to time-series (see Table below).

[95] Some authors present the trend and cycle as a single component, the trend-cycle.
[96] Loess – Locally weighted regression and scatterplot smoothing

Table 5: Type of data transformation

Method		Properties	When to use
Seasonal adjustment	Multiple	Removes a constant seasonal pattern from a series	When the seasonal component needs to be separated out to fit with a non-seasonal model (e.g. regression)
Differencing (random walk)	$y_t - y_{t-1}$	Converts *values* to *changes*. The differenced series is the change between consecutive observations in the original series	To transform a non-stationary process to a stationary process
Differencing (second order)	$(y_t - y_{t-1}) - (y_t - y_{t-2})$	Converts *values* to *changes*. The second-order differenced series is the change between consecutive observations in the first-order series, i.e. the random walk	Occasionally the differenced data will not appear to be stationary and it may be necessary to difference the data a second time to obtain a stationary series
Differencing (seasonal)	$y_t - y_{t-m}$	Converts *values* to *seasonal changes*. The differenced series is the change between sequential observations of the same season	If the seasonal pattern is consistent and it is required to be maintained in long term forecasts
Log	$\ln(x)$ $\log_{10}(x)$	Converts multiplicative patterns to additive patterns and/or linearizes exponential growths	Helps stabilize the variance of a time series when data distribution is positive and highly skewed or when variables are multiplicatively related

6.3.7. Forecasting

As presented above, the purpose of conducting time-series analyses includes forecasting values. The term forecasting suggests a prediction of values or state of a variable in a future time, in fact, when used as a monitoring tool, which is likely to be the main use in terms of regulatory activities, ex-post forecasting is more relevant.

Ex-post forecasting is the prediction of values that have already been observed, forecasted and observed values can then be compared. Consider for instance performing ex-post forecasting of the number of case reports of an adverse reaction, this could be compared to the observed value in a time interval, if the observed value is higher than the upper estimate, it may indicate a new safety concern.

Forecasting methods are typically roughly divided into time-series methods and causal methods, where time-series methods are univariate methods and casual methods are regression models, which can be simple or multiple.

In fact, the same methods can be used for all of the aforementioned purposes. For instance, regression models may help explain the behaviour of a time series as well as predict it in the future, provided that the assumptions are not violated.

6.3.8. Change point detection

Change point detection is related to forecasting. A *change point* represents a transition between different states in the process that generates a time series. In other words, a transition between one deterministic series to either a new deterministic pattern or a random one.

As the change point is a point in the time series where its properties change, it can be considered as a hypothesis testing between two alternatives, the null hypothesis H_0: "No change occurs" and the alternative hypothesis H_1: "A change occurs".

For instance, a new contraindication in a medicinal product in patients with chronic renal failure should have an effect on the deterministic pattern of the prescription of the product in renal failure patients, reducing it, which in turn, should have an impact on the occurrence of the adverse reaction. If a change is observed, then one can assume that the regulatory action had a deterministic effect.

Change point detection is relatively new in the pharmaceutical regulatory context but has been widely used in other scientific disciplines, which have led to heterogeneity in terminology. *Change points* are sometimes referred to as *changepoints* or *breakpoints*, whereas *Change Point Analysis* is sometimes referred to as *Joinpoint Analysis*. Joinpoint however, is also the name of statistical software developed by the US National Cancer Institute that is presented as using Joinpoint regression, which is a segmented regression, a type of change point method.

Moreover, the term *interrupted time-series analysis* (ITS) is often referred to in healthcare as a specific type of model, whereas would be more adequately referred to as a subtype of change point analysis, where the change point is hypothesised to be at the time of a specific intervention. The other form of change point analysis is change point detection (CPD), where the change points are data-driven rather than predetermined. In fact, CPD can be considered as an iterative testing of the occurrence of change at all observations of a time series. Segmented regression is one model that can be used in ITS, but others can also be used.

Considering the variety of terms and that *join point* is also a term in computer science, it seems better to refer to the wider activity as *Change Point Analysis* and the changes to be named *change points*. Change point analysis can happen in two forms; *ITS* to be used where a hypothetical change point is tested and *change point detection* (CPD) where the data drives the change point detection.

Noticeably, most research using ITS acknowledges a limitation with regards to defining when the intervention began. Take the example above, the point at which a new contraindication takes effect is not the moment of the recommendation and could be months from then. It may be wiser then to use both ITS to test the defined time point, but also CPD, which, unconstrained by a priori hypothesis, may provide a better insight as to when a change manifested in the time series.

Whatever the type of CPS, the statistical properties that are tested for changes are normally the mean, the variance, the regression slope, or a combination of mean and variance. As can be deduced, where the mean and variance are tested, the assumption of stationarity is required, such that the change is where stationarity is breached. Conversely, for regression slopes, the assumption of no correlation of errors is assumed. Where these assumptions are violated, data transformation is required.

6.3.9. Time-series models

There are several time-series models, for a wide variety of these the mathematical concepts behind them are presented in the supporting literature, as the discussion of the mathematical concepts behind the models is outside the scope of this document.

Some models are simple and rarely used in practice such as the average method, the naïve method, the seasonal naïve method, the drift method, moving average model and the autoregressive model.

The most sophisticated and widely used approaches are ARMA models (autoregressive moving average), ARIMA models (autoregressive integrated moving average) and Exponential smoothing. The difference between the ARMA and ARIMA models is that in the ARIMA model differencing is applied.

ARIMA models

ARIMA models can be considered a general class of models that includes random walk, random trend, seasonal and non-seasonal exponential smoothing and autoregressive models, where forecasts for the stationary/stationarised dependent variable are a linear function of lags of the dependent variable and/or lags of the errors.

Non-seasonal ARIMA models are typically denoted as ARIMA(*p,d,q*), where p is the number of time lags of the autoregressive model, d is the degree of differencing (e.g. first order, second order) and q the order of the moving average model. Seasonal ARIMA models, or SARIMA, are typically denoted as ARIMA(*p,d,q*)(*P,D,Q*)$_{m,}$ where *m* is the number of periods in a season.

ARIMA models can be used when data are plentiful, with more than 3 seasons, and can be made approximately stationary by differencing and other mathematical transformations.

While these models are typically run as univariate models, it is possible to include covariates in the ARIMA models, resulting in what is called ARIMAX models.

Exponential smoothing

Exponential smoothing is presented here as distinct to ARIMA, however technically it can be considered a type of ARIMA. In these models, exponential functions are used to assign weights over time. A simple exponential smoothing would then be the same as an ARIMA (0,1,1) model without a constant.

These models can be used when the data are non-seasonal or can be de-seasonalised. This method is suitable for forecasting data with no clear trend or seasonal pattern.

Regression models

Regression models are causal models that can be used to describe or forecast time-series. The most basic of these models is the linear trend, where a regression of the observed value on the time index is performed. However, this is a fairly unreliable model for explaining the behaviour of the time series or for forecasting and should be reserved for situations where very few data points and no obvious pattern, other than a trend, can be seen. Where seasonal variations are evident and there are enough data, it can be used with seasonal adjustment.

Multiple regression can be used when data are correlated with other explanatory variables. In these cases, the key is to choose the correct predictor variables and the adequate transformations of the variables to justify the assumption of a linear model. Multiple regression models are better placed to explain time-series behaviours rather than predicting a future event as these require the explanatory variables to be known.

Segmented regression or piecewise regression is, as seen, a specific type of regression often used to perform ITS analysis. It differs from the linear trend in that model splits the data at the point of an intervention, sometimes known as the knot location resulting in two separate linear functions. The intercept and slope of the two are then compared to assess the difference.

In the regulatory context, count data will be mostly used, and in such cases, the preferred distribution to be used is Poisson, or where overdispersion may be present, the negative-binomial distribution.

Other available models to adjust for count data with excess zero counts, include the zero-inflated Poisson and zero-inflated negative binomial models.

Models to detect changes

As addressed above, there are also a variety of models to detect change points. Cumulative sum charts (CUSUM), Binary segmentation (BinSeg), At most one change (AMOC), Pruned Exact Linear Time (PELT). These detect changes in mean, variance or mean and variance. Some are recent and increasingly used, such as the PELT method, however all these can be sensitive to minor changes in function arguments, such as the minimum sequence of time for a change to be considered.

Granger causality

Most time series discussed so far assume that the data are univariate. There is a particular type of scenario where one time series can be used to forecast another. Consider for instance the date of occurrence of an adverse drug reaction and the date when it was reported. For the same adverse reaction this interval should be a function of the type of reaction (acute or chronic or insidious) and of the reporting requirements such that there will be a relationship between the date when it occurs and the date when it is reported. In case significant media attention occurred, the reporting of historical cases might void the granger causality, suggesting an artificial increase in the reporting.