Clusters of Excellence Discussion Paper

23 May 2022

Index

Executive Summary

This document is the joint work of a group of dedicated people across some of the European Medicines Regulatory Network (EMRN) agencies to embed data analytics into the everyday work of the EMRN. Six main building blocks for the creation of Clusters of Excellence (CoE) in the EMRN form the framework for this discussion paper: data access, legal aspects, capabilities, infrastructure, methods development, and artificial intelligence.

Representatives from interested agencies have worked together to answer three fundamental questions relating to each of the above-mentioned building blocks: What is the current state of play? What are the challenges we face? What are the opportunities for collaboration?

The current state of play is enriched with concrete use cases from select agencies which can be used as a blueprint for the improvement of the data analytics capabilities in other agencies or the establishment of a data analytics capability in an agency that does not yet have such a function. The overview of identified challenges by the agencies can be used by the EMRN to identify areas of future action.

In the discussion process, four initial clusters of excellence emerged: a Cluster on AI (BfArM, DKMA, PEI, MPA), a Cluster on High-Performance Computing (AEMPS, BfArM, DKMA, PEI, MPA), a Cluster on Real World Data (AEMPS, BfArM, DKMA, INFARMED, MEB) and a Cluster on Patient-Level Data Analysis and CDISC (DKMA and MEB). To expand the number of clusters and the number of members of each cluster this discussion paper contains a range of suggestions for collaboration that can be summarized into the following actions:

- Establishment of knowledge sharing forums for the exchange of insights on the relevant building blocks;
- Standardizing legal agreements for data access to lower the cost of gaining data access;
- Create a portfolio of use cases as inspiration for the entire EMRN;
- Harmonization of terminology by creating an official glossary;
- Development of best practices and standards within regulatory data science, data management, and required software;
- Collaboration on the training of new talent by creating a CoE inspirational curriculum;
- Create a European data catalog of available data and share reference data for the training of algorithms;
- Collaboration on harmonizing metadata & platforms between agencies as well as core variables when developing registries;
- Establish a process for how to help agencies with limited data analytics capabilities to accelerate their development;
- Collaborative effort to ensure future-proof regulation;
- Establishing an EMRN-forum for the sharing of results and experiences in projects under Horizon Europe Tools 11-02.

By implementing the above-suggested actions, we will move the EMRN as a whole closer to archiving excellence in health data analytics and create the necessary foundation for the future integration of our national agencies into clusters of excellence. The recommendations from this paper will be fed into the workplan review currently underway at the Big Data Steering Group.

Foreword

This document is the joint work of a group of dedicated people across some the European Medicines Regulatory Network (EMRN) agencies with the aim to embed data analytics into the everyday work of the EMRN.

During the work of the Big Data Task Force it was clear that the use of analytics in the national competent authorities was taking up speed. During 2021 EMRN data experts, as an EMRN cluster of excellence of data analytics across Europe met together to discuss our progress to date, our challenges and how we could collaborate to advance the use of data analytics in our work.

This document is a snapshot of the current state and will change constantly. This report by a group of experts from some agencies from the EMRN was endorsed as an expert report by the Big Data Steering Group (BDSG) in April 2022. The observations and suggestions in this paper will be fed into the 2022 review of the BDSG work plan (due for adoption June 2022) and the review of the BDSG mandate (due by early 2023). Where possible existing bodies and initiatives will be leveraged to take forward any key recommendations.

It is the intention to expand the group over the years to come and through our collaboration expand on the agency and national capabilities in analytics and use this in both our national and EU work.

Introduction

Based on brainstorming sessions in the last quarter of 2021 conducted by the Clusters of Excellence (CoE) sub-group of the BDSG comprising representation of National Competent Authorities (NCAs) from DE, DK, ES, NL, PT and SE, a range of statements relating to six main building blocks for creating Clusters of Excellence in the EMRN were articulated. The six building blocks have been defined as: data access, legal aspects, capabilities, infrastructure, methods development and artificial intelligence.

In this discussion paper representatives from interested agencies have worked together to answer three fundamental questions relating to each of the above-mentioned building blocks: What is the current state of play? What are the challenges we face? What are the opportunities for collaboration?

The following sections summarise sessions, with the aim to stimulate more discussion with other EMRN Agencies and to develop concrete areas for collaboration.

Data Access

Current state

Most agencies have access to several data sources, but there are differences between the agencies.

Multiple agencies contributing to this paper have experience with different ways of gaining access to data. Some agencies have made efforts to link registry data, others have taken the initiative to create their own registries as well as gained access to patient-level data on prescription, claims and reimbursement data as well as population based pharmacoepidemiology databases.

Below you can find a collection of examples from different contributing agencies regarding data access.

Portuguese National Cancer Registry
In 2017, a National law (Lei n.º 53/2017 de 14 de Julho) was published that centralises the regional cancer registries in a central platform. Article 2 of the mentioned law states that this national registry also aims to monitor therapeutic effectiveness in collaboration with INFARMED. In article 5 of the law is mentioned that specific data can be collected in the situations requested by INFARMED and should be collected within indicated timelines.

The National Cancer Registry has been used to monitor effectiveness in the reimbursed indications of specific medicines, with clinical uncertainty and high economic impact.

An example of this use is presented in the article "Monitoring real-life utilisation of pembrolizumab in advanced melanoma using the Portuguese National Cancer Registry.", published in Pharmacoepidemiol Drug Saf. 2021 (doi: 10.1002/pds.5163).

The registry is used to monitor pembrolizumab, nivolumab, palbociclib and the reimbursed CAR T cells. INFARMED also develops its own registries (hepatitis C, Spinal Muscular Atrophy and Lysossomal Storage Diseases).

Access to data via collaboration in the Netherlands
MEB has different collaborations with registries (such as NIVEL, IKNL, DICA). However, MEB does not have access directly to the data. Recently, the Health-RI initiative was funded by the government, which intends to link all registry data together, and this is a very interesting initiative for the MEB.

Linking data sources in Germany
PEI is part of a project to link patient data from health insurances and vaccination status from the Robert Koch Institute to analyse the side-effects of vaccinations. Data pseudonymisation, encrypted and transferred via Bundesdruckerei (Federal Printer). BfArM has access to public health insurance data, collaboration with public health insurance. The BfArM is establishing the new health data lab which provides access to health insurance data of all persons in Germany with statutory health insurance.

Primary care database in Spain
AEMPS uses medical records from the primary care to create a population based pharmacoepidemiology database in order to generate scientific evidence to support regulatory decisions in different contexts. Database created and maintained by the AEMPS: The BIFAP Program: http://bifap.aemps.es/index_EN.html . AEMPS also has access to a wide ranger of data through collaboration as the Netherlands.

Nationwide registries in Denmark

At the DKMA a wide range of nationwide registries covering medication, diagnoses and diseases are made available through secure computer environment at the Danish Health Data Authority, the national Danish data permit authority.

Challenges

An overview of the current key challenges regarding data access faced by the contributing agencies is provided below.

Governance structure

The lack of a clear governance structure is currently a challenge. A clear governance structure coherent with the purposes of the EMRN would facilitate access to national datasets. The establishment of the DARWIN EU coordination centre could provide an opportunity to create governance modalities as some of the data access issues will be common.

Data in the hands of third parties (regions, hospitals) is not always easily accessible and the lack of that data could provide some risks to the validity and robustness of the analyses performed.

Lack of data standards and integration and linkage of data

There is currently a lack of harmonisation of data variables in order to facilitate joint analysis between different entities and countries.

A unique identifier that links the different EHR/registries at national level is needed and, in many places, missing.

Availability and differences in entries in EHRs (structured vs unstructured, granularity etc.) in the different countries may impose limits on possible analysis of such data.

Lack of sharing overview of European data sources

Limited knowledge about the different national registries and variables and how they can be used. Which agencies do already have access to certain datasets & registries?

In order to support infrastructure for data access, precise meta data knowledge of national health care systems and electronic registries of prescribing, reimbursing medicines should be identified.

In could be helpful to create an EMA metadata catalogue. It would be nessecary to create a process and/or incentive for maintaining the catalogue as to ensure that it is always updated.

Incentives for data collection

Need to improve incentives and establish consent/permission to have access for regulatory authorities for the use of the data.

Other remarks

Lack of relevant training or education required to access the data, aka technical, secure and legal requirements for compliant data access.

Inefficient data collection due to lack of linking. We should avoid duplication of registries and databases and extract data directly from the source. This would avoid overburdening doctors with administrative requirements. Risk of affecting quality and completeness of data

Collaboration

Based on the current situation examples and the challenges, and the experiences across agencies, we identified a number of areas for collaboration to improve the current state as suggested below:

Data Standards

- Promote harmonised core variables when developing registries.
- Support establishing standards for data linking.
- Promote and apply common data models to enable a common analysis of data across countries. In alignment with the EMRN "Data Standardisation Strategy" .
- Explore DARWIN EU as an engine for the promotion of common data standards.

Knowledge sharing forums

- Promote a structured dialogue and knowledge sharing between different agencies working on the life cycle of medicines. The info needed for medicines regulatory purposes could be used also to HTA reassessment, with only some adjustments.
- Share knowledge on ways of accessing data and how to overcome some constraints, including legal aspects.
- Share knowledge, e.g. on methods; topics from all areas within a lifecycle of a product, as also seen in practice.
- Create an analytics group (or groups) which is/are able to deal with operational issues from a network perspective.
- Sharing of metadata and potentially increase uniformity on metadata.

Access to data in relation to EMA procedures

It is assumed that EMA data access activities are currently mostly for support of EMA committees. Use cases addressing research questions from EMRN agencies should also be supported by central data access structures in the future. Example: Collaborate on Pharmacovigilance Risk Assessment Committee (PRAC) referrals: use of relevant data where available.

Related with UNICOM, there are opportunities to leverage the enormous amount of data that regulatory authorities hold.

Other remarks

International platforms to access data (e.g. DARWIN EU) vs. 'personal' access to datasets: not an either/or but both depends on the research questions, cross national, national, regional etc.

To develop a kind of best practice guidance to sign agreements with data providers, share templates and agreements for inspiration.

Important to have access to multiple countries/regions, as there might be country/region-specific differences. In the rapid data analytics project for the PRAC also different databases/countries are consulted.

Legal Aspects[1]

Current state

Currently, many of the regulatory agencies have access to pseudonymised data and several have a system of data access governance in place. Also, agencies work within the use of anonymous data (aggregated) if possible, where GDPR does not apply. Still, the amount of paperwork required for receiving anonymised data is high and the process is long.

With regards to different types of data there are several agencies in the EMRN currently looking into working and receiving applications in CDISC-format. However, there might be added potential legal issues if the informed consent forms (ICF's) put restrictions on use. Furthermore, some agencies struggle with the lack of clarity on the legal basis for requesting CDISC-data from the pharmaceutical companies.

Examples

Currently several agencies are challenged in the field of data access. One approach to pave the way for the agencies and the cluster of excellence is to incorporate incentives for data access in national legislation as in e.g. Portugal (the law that created the National cancer registry defines that the data can be used by INFARMED to monitor effectiveness of medicines and medical devices - In addition, for two disease registries developed by INFARMED to monitor effectiveness (Hepatitis C and Spinal Muscular Atrophy), the financing of drug treatments for these diseases in the hospitals is based on the registry.). It is important to flag to the government and legislative bodies that incentives should include the regulatory agencies and ensure data access through legislative measures. The analysis capacity and benefits for the public if incentives are incorporated are great. The incentives for data access should include health data in relation to both pharmaceuticals and medical devices, and enable the regulatory agencies to make relevant analysis for the benefit of the European public.

Challenges

With regard to the challenges in the current European legal landscape, it is evident that there are multiple aspects that are worth investing with more time and effort in order to accentuate the process of creating a CoE. In the following these subjects will be explored further.

Legal requirements for compliant data access

There are multiple initiatives across Europe in gathering various forms of health data. Currently every European country sets its own regulatory framework for the processing of health data. Healthcare is a national competence and the data obtained from healthcare are not shared cross-border as often as research data. There is therefore a significant challenge in guiding national and EU agencies in compliant access to national health data. This has been addressed in the domain of genomic data in a joint declaration by 21 EU member states to deliver cross-border access to genomic data by the end of 2022. (see e.g. Leveraging European infrastructures to access 1 million human genomes by 2022 | Nature Reviews Genetics)

[1] Add reference to EHDS legal proposal when published

Legal basis for requesting data access

As capabilities and potential of processing raw data increases in national authorities, it is likely that there will be increased interest in access to more data from industry. Experiments in this field are currently conducted, and could help pave the way for a more efficient and expedient process for marketing applications for medicinal products. If this is a model to be elaborated more, it is necessary to elaborate on the legal framework underpinning access to raw data: when and on what legal grounds may regulatory authorities require data from e.g. industry?

Incentives to share data

To share data may be extremely cumbersome for the provider e.g. a researcher or a company. Incentives or compensation mechanisms for such efforts could help increase the amount of data sharing. It should however be considered further how such arrangement could be designed, its feasibility also in terms of administration, the proper level of implementation (national or EU, industry or general).

Implementation and interpretation of EU regulation

Different legislative initiatives have been launched in order to provide a legal framework that could accommodate data sharing, e.g. the GDPR, Open data directive and Data Governance Act (still to be approved). Despite that these initiatives have been put in place to i.e. increase harmonisation between member states, national discrepancies in the interpretation of the concepts and requirements continues to subsist.

Standardising legal agreements

In line with the ideas of further igniting the cross-border collaboration and sharing of data, there is also a wish for "standardising" legal agreements (e.g. templates) for data access and further. In connection with this it is also necessary to look further into the protocols for the use of data stored in the cloud, in order to streamline the process in the EU. An important aspect that we have to deal with are contracts with external entities (access to data and confidentiality issues). Third-party use is different from direct use. When we look into standardising legal agreements (e.g. User agreements, License agreements and further) it is important to consider the discussion of what the European standard should be, and where to set the bar.

Collaboration

Knowledge sharing

A large part of the legal body that puts constraints on data analytics (now or in the near future) is European in scope e.g. GDPR, future harmonised rules for the use of Artificial Intelligence etc. For this reason, it would be helpful to collaborate on sharing knowledge about the interpretations and limitations of the relevant regulation in order to 1) navigate the regulation most efficiently 2) identify sub-optimal regulation that could be sought to be amended. Better knowledge of the legal landscape would also make it easier for the EMRN agencies to input into the discussions of the regulation surrounding the EHDS.

The knowledge sharing could take place through a legal forum with representatives from all interested EMRN agencies.

Training/recruitment

In order to establish the necessary legal expertise, it would be helpful to collaborate on developing training material e.g. an overview of which legal topics should be covered during on-boarding of newly recruited

legal resources. The curriculum could include topics like legal requirements for compliant data access, anonymisation, privacy protection, data ethics etc.

Coordinated effort to ensure future-proof regulation

A concerted effort to ensure that new regulation e.g. in relation to the EHDS is facilitates the work of medicines regulation could be an area of collaboration. One option would be to establish a European group for the coordination/discussion of current or new regulation relating to areas of interest for the CoE. The group would enable the EMRN agencies to give informed input via the relevant national channels to the European legislative process and thus help make new regulation future-proof.

Capabilities

Current state

Below is an overview of the current state regarding capabilities in the contributing agencies. This theme overlaps with the theme on new methods, but here we focus on the human aspects and the knowledge and competencies that make up the agencies' capabilities.

Competencies in the agencies

Currently, the agencies already covers a broad spectrum of expertise, research areas and methods. Within the organisations, these capabilities are mostly centred in small teams or sole experts. This varies from a team of methodological experts or epidemiologists to a single data scientist.

Competencies currently being evolved

Within the MEB and DKMA, the focus is mostly on evolving the data analysis skills by building a dossier that includes researching Individual Patient Data (IPD – formatted in CDISC SDTM and ADaM datasets), gaining knowledge regarding data standards (e.g. CDISC SEND at MEB) and building knowledge around the use of real-world evidence (RWE).

MEB to stimulate the development of capabilities.
The MEB has published a Science policy for the years 2020-2024. The science policy aims to advance and optimise systems & processes and to innovate on topics related to the field of medicines regulation. The policy covers 8 themes:

1. Replacement, reduction and refinement of animal tests (3Rs)
2. Advanced Therapy Medicinal Products (ATMPs)
3. Data-driven assessment
4. Personalised medicine & biomarkers
5. Medical devices
6. Generics
7. Medicines used better
8. Safety and effectiveness after authorisation

In addition to the evolving processes of the MEB, the Paul Ehrlich Institute (PEI) is researching how to evaluate AI algorithms from a regulatory perspective by asking questions on how to ensure the quality of a statistical model through validation when it is used for the production of biomedicines. In the research division of the BfArM several interdisciplinary research units, including pharmacoepidemiology and biostatistics, focus on regulatory science topics related to authorisation and improvement of safety of medicinal products (https://www.bfarm.de/EN/BfArM/Tasks/Research/_node.html). There is also regulatory research on the risk identification and assessment of medical devices. The research units apply a broad spectrum of methods in data analytics, including AI-based algorithms. Scientists closely collaborate with leading national, European and international research institutions.

As stated above, the regulatory network is exploring and gaining knowledge on various capabilities, mainly with two goals in mind: 1) aim to research the current status of the capabilities and 2) aim to improve the capabilities. These pilots and projects are mostly focused on building up competence in the field of data

management and data analyses, where software seems to play a bigger role than hardware. For some organisations however, time and support seem to be inhibiting factors.

External collaboration

The agencies are involved in various collaborations with academia and other regulatory agencies. For example, the data mining capabilities are evolving through a collaboration with the Utrecht University and the University of Copenhagen. Other examples are Regulatory Science Network Netherlands (RSNN) and Copenhagen Centre for Regulatory Science (CORS at University of Copenhagen) platforms to advance and improve the regulatory system together with stakeholders from the industry, academia and government bodies. Another example is the close collaboration of the BfArM with the University of Bonn, including joint research activities and teaching (e.g. study course Master of Drug Regulatory Affairs).

Within the regulatory network, the call to research the use of data standards is also explored through a collaboration between agencies in a pilot which aims to gather more knowledge about the use of one of CDISC's data standards. This involves a number of capabilities to be established between statisticians, data scientists, IT and legal across the known capabilities placed with medical assessors.

However, seconded roles at committees and joint projects also play a major role in gaining knowledge and expertise on capabilities as well. One of the examples is the role of the agencies in the BDSG, which raises awareness at the agencies to become more data driven.

These collaborations ensure knowledge transfer within the network and thereby indirectly prevent the network from doing 'redundant work': research, project or pilots that have already been done before at other organisations.

Example

In order to ensure the availability of the right competencies the contributing agencies have taken a range of measures to attract new talent as well as developing the personnel already working in the agencies. Please see the box below for an illustrative example from the MEB.

Investing in new talent

The MEB has invested in a PhD-project that intends to build a dossier by identifying factors that contribute to the (non-)acceptability of the use of RWE as a supplement to – or substitute for – evidence from RCTs in regulatory decision making. The project intends to:

1) build theoretical knowledge which can be applied in real cases by the EMA and National Competent Authorities.

2) serve as a catalyst for a lively discourse around the topic of RWE.

Challenges

Below is an overview of the current challenges regarding capabilities faced by the contributing agencies.

Training needs

The use of data standards or using data science to answer regulators' questions requires special training for assessors to become experienced with these novel technologies and to apply them efficiently so it becomes a time- and workload-saving asset.

Challenges with attraction and retention

Capabilities are often tied to personnel, hence developing or improving capabilities requires training of current employees and attracting new experts. Agencies especially need new experts in the fields of epidemiology, statistics, data science and artificial intelligence. To attract these experts, agencies have to compete with the industry.

Financial challenges due to lack of capabilities

Hiring experts requires financial investments, which may be costly or sometimes not feasible because of the competition with the industry. Besides attracting expertise, training and development of capabilities may be time-intensive and can therefore become costly.

Other remarks

The regulatory network addressed some remarks to the topics mentioned above:

- The regulatory network is waiting for the scope of the clusters of excellence to be defined. Once the scope has been defined, organisations may adjust their vision on the development of capabilities accordingly;
- A protocol or framework should be established which describes how to work with organisations with less (or more) capabilities to avoid miscommunication and inefficient knowledge transfer that guarantees the equal knowledge transfer between all agencies and thus does not distinguish between frontrunners or leading roles like rapporteurships;
- Manage capabilities: Create an overview of which EMA and EMRN agencies processes that could utilise or be supported by data analysis.

Collaboration

Below is an overview of the opportunities for collaboration regarding capabilities identified by the contributing agencies.

Knowledge sharing forum

A place for knowledge transfer should be created where the regulatory network is able to share knowledge through workshops, educational material and interactive discussions (e.g. forums) on the topics of data science, methodology and required software. This place for knowledge sharing should also facilitate the sharing of 'physical' capabilities such as code to perform analyses (e.g. R scripts).

Besides knowledge transfer, the forum can also be used to:

- Define protocols and frameworks which can be used within the regulatory network;
- Discuss important topics for future pilots and projects;
- Share expertise on the use of national healthcare data.

Best practice and guidelines

Besides the place for knowledge transfer described above, a white paper containing best practices could be established on the topics of regulatory data science, data management and required software. This white paper could be created through collaboration between the agencies.

Training

As a supplement to the best practice paper and the knowledge forum, training could be provided to stimulate the development of capabilities within the regulatory network. Agencies that have a lot of experience with certain capabilities could train other agencies. Multi-agency teams could be established to tackle certain

topics, workshops can be given and already existing educational material should be utilised and extended (e.g. the existing biostatistics curriculum at the EU-NTC).

For example, the BDSG is developing multiple training curricula for the EU Network Training Centrum (EU-NTC). The topics covered by the curricula are Pharmacoepidemiology/real-world evidence, Biostatistics & Clinical Trial Methodology and Data Literacy. These have been developed as part of the 4th recommendation by the BDSG *"Develop EU Network skills in Big Data"*.

Infrastructure

Current state

Most contributing agencies have made significant progress in improving their infrastructure in order to be able to cope with the ever-increasing amount of data. Many agencies have access to high-performance computing (HPC) capabilities for the storage and analysis of large and complex datasets. The access to HPC can either be through own investment or collaboration with external providers e.g. research institutions or other governmental agencies. Moreover, the agencies have invested in software and analytical tools as well as the necessary competencies, in order to be able to take full advantage of the advanced hardware infrastructure. Please see the box below for a more detailed example from the MPA, BfArM and PEI – similar examples could be shown for other agencies:

High-performing infrastructure in Sweden
The Swedish Medical Products Agency (MPA) has a well-developed infrastructure for protected high-volume data storage, and well-established protocols and hardware for acquiring, curating and analyzing national registry data. In-house modeling has previously been restricted to the field of pharmacometrics but is now being expanded to include the full range of data science applications including machine learning.

A GPU-powered (RTX3090) AI server hosting JupyterLab/Keras/TensorFlow services along with Python data science libraries is available for in-house model development and evaluation, along with servers for model deployment in the internal production environment. For high-performance computing projects, we have access to the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX).

The MPA is in a capability and competence building phase, currently recruiting software developers and engineers as well as a PhD student within the field of data science in cooperation with a Swedish university.

BfArMs high-performance computing infrastructure
The Federal Institute for Drugs and Medical Devices (BfArM) is expanding AI-computing infrastructure in different areas, including licensing, Health Data Lab and research. For example, the IT cluster of the research division consists, among others, of two IBM® Power System™ servers (AC922) for High-Performance Computing and AI computations, which operate as a multitenant architecture in a shared environment optimized for deep learning analytics using the IBM Watson Machine Learning Accelerator.

Next to established statistical analytic tools (e.g., R, SAS) to perform queries on data and extract information on drug utilization and diagnoses within the healthcare system, modern approaches including machine learning and respective software (e.g. python) are being used.

For instance, by applying artificial neuronal networks for pharmacoepidemiological analysis, or via artificial intelligence assisted text mining in signal detection, respectively. Moreover, BfArM is establishing a new infrastructure to provide access to health care data, with the Health Data Lab. The Health Data Lab will provide permission as well as access to data of all persons with statutory health insurance (more than 70 million people) in Germany via a secure analysis platform.

Health Data Lab experts are also active in European projects such as TEHDAS. In scope of interoperability, BfArM is the National Release Center (NRC) for SNOMED CT. Another important

collaboration partner of the BfArM is the gematik which provides the telematic infrastructure for electronic health records which will be also accessible at the Health Data Lab.

PEIs infrastructure
The Paul Ehrlich Institute (PEI) is planning a cluster of four 64 core compute nodes for early 2022. Each node will be enhanced with A100 tensor core NVIDIA GPUs to facilitate machine learning analyses. The institute maintains and develops internal user and programmatic interfaces facilitating pipeline automation, and data access and visualisation. Live dashboards for the visualisation of time-critical data are planned for vigilance projects

Challenges

Hardware

- Larger datasets will require considerable computer power for analyses.
- For hardware and software to be able to process a big amount of data in due time, machine time matters.
- Risk management is needed with the use of IT infrastructure.

Infrastructure to collaborate with data between agencies

- Storage: Where and how do we store the data?
- Lacking infrastructure for data pipeline from party to party meaning data can be out-of-date.
- IT infrastructure: currently there is no 'network' of IT infrastructures between agencies.
- Ideally, we should get some data from all/most MSs for some fundamental issues. How to get this from some agencies that do not have any prior experience on this?
- Currently only very limited options for cross-border co-operation regarding data access are established.

Other remarks

- Clear infrastructure / process description - who has which data, and where can it be used for, potential costs (and if so, is there money for a pilot for instance)?
- Important to have a fast answer to be able to use it in regulatory procedures.
- Will analyses be run on own machines or will this be outsourced?
- Lack of access to data, standardisation.

Collaboration

Catalogue of infrastructure options

- Potential of cloud services shared among agencies or is access to IT-infrastructure of other agencies possible? If so, how?
- Harmonising metadata & platforms between agencies.
- Tutoring of non-expert agencies by contributing agencies as an exercise to practice the BDSG. recommendations and help agencies gain the capabilities to be part of a Cluster of Excellence.

Data access and use cases

- We could learn from how data access is arranged in the individual agencies.
- Sharing reports or outcomes within the network of which tools and software are used for data analysis (e.g. for using FHIR, OMOP and CDISC formats).

- Workshops between agencies, presenting some studies and the process to obtain data and analyse it.

Development of new methods

Current state

The contributing agencies agreed that the development of new methods is an imperative to be able to unlock the potential of the vast amounts of data that is being generated within the sphere of medicines and medical devices. Multiple agencies are currently working on expanding their portfolio of advanced methods with data analytics. The agencies strategy is two-pronged: 1) building knowledge of existing methods and designs and 2) developing new methods for data analytics e.g. within AI and RWE. For a more detailed example on PEI's efforts to develop new methods for AI please see the example box below.

Challenges

Resources and capabilities

- Resources are limited (in terms of experts and machines).
- Lack of internal capabilities – reinforce collaboration with academia.
- Limited time available, given costs, but also current workload of assessment within the system.
- "Brain drain" due to departure of employees involved in the development of new methods.

Overview and collaboration

- Collaboration with industry and data transparency.
- Communication with regulatory network on what is currently being developed.
- To be up-to-date on innovation in this field and how to get in contact with external expertise.

Access to data and standardisation of data

- Access to structured/complete data.
- Standardisation of methodology among regulatory agencies (i.e. guidelines e.g.).
- Data sources have variable quality and data content may not be suitable (covariates missing).

Other remarks

- Advantages of complex methods are not clear and have to be demonstrated.
- Evaluation of functionality/implementation of new methods.

Collaboration

Knowledge sharing and knowledge sharing fora

- Discuss evaluation of novel methods.
- Identify topics: develop a list if "use cases", or scenarios, since many in common across agencies.
- Working groups on identified challenges/new areas.
- Sharing expert knowledge (e.g. complex trials including use of RWE).
- Good communication between agencies, so developments can strengthen each other or complement, and not overlap too much.
- Share practice and further down the road create best-practices among European regulators for analysis/inspection of machine learning (other AI based models) and including devices with machine learning modules.
- Create a forum for agencies to discuss "cases" which challenges regulatory framework.
- Create a forum/platform for external experts to join and contribute to development of good practices.
- Create a way to share the analysis that are being done. Workshops where agencies could present their studies.

- Experts from different agencies could work together in teams to develop new methodology jointly.
- Have a central priority list of needs for development.

Resources and education

- Courses that could develop more focused capabilities to internal staff of agencies.

Sharing data and data standards

- Cooperation on data standards and format necessary (to enable data exchange).
- Equal requirements to data quality and access.
- Sharing of datasets as “reference data” for algorithms development.

Artificial Intelligence (AI)

Artificial intelligence (AI) is quickly becoming commonplace in the medical sector and is being used for many different purposes. AI and machine learning have been reported to be used in the manufacture of personalised medicines as a companion diagnostic for immunotherapy. In such an application, machine learning models are used to predict neo-epitopes that are specific to a patients' cancer cells and can be used to trigger a targeted immune reaction against the cancer. AI and machine learning have also been reported in the development of CRISPR-Cas based medical products. Machine learning models have been developed to predict the likelihood and frequency of off-target editing for a given guide sequence. In this manner, guide sequences can be chosen that are least likely to introduce mutations into coding genes and other functional parts of the genome. Not only is AI being used in the development and manufacture of medicinal products, but also in regulatory processes. Detailed in the examples section below is a description of how natural language processing is used in pharmacovigilance by the Swedish medical products agency. In this section we seek to assess the current AI-related activities and capabilities within the CoE, what challenges AI presents the agencies involved and how we may collaborate together on overcoming these challenges.

One area of AI application in the medical sector that we do not seek to address, is the use of AI in medical devices including applications such as expert systems and machine learning models used for patient adjudication.

Current state

Currently, the regulatory network as a whole has limited AI capabilities and the ability to develop them. The need for such capabilities is even debated, with some agencies showing limited interest in developing in this direction. Nevertheless, other agencies are interested in AI for two major purposes. 1) The use and development of AI to support regulatory activities within the organisation. 2) The validation and regulation of AI used in the development of medicinal products.

Resources

Agencies in the EMRN demonstrate a highly heterogeneous landscape of resources and expertise when dealing with AI. Some agencies do not see the use of AIs as part of their main mission or express little interest in developing AI capabilities. The availability of personnel and infrastructure limits others. Only a few have ongoing AI related projects. In general, agencies with active AI projects have small teams of data scientists, biostatisticians and bioinformaticians and limited computational infrastructure.

Methods and validation

No participating agencies reported the use of AI to assist current regulatory activities. However, there are ongoing research projects covering a broad range of topics investigating this possibility. There are also research projects investigating new data sources that applicants may use such as the analysis of real-world data (RWD) as a supplement for evidence from randomised controlled trials. Other research projects investigate the use of AI in incidence reporting or the development of a "regulatory brain" that uses natural language processing to assist regulators in the evaluation of submissions through the analysis of historical data.

The current validation of AI used in the development of medicinal products is restricted to assessing the suitability of machine learning algorithms for addressing the posed problem, the generalisability and biases of the training data set, the predictive performance of machine learning models through the use of performance measures such as sensitivity and specificity, evaluation of model training logs, and the stability of the environment in which the model is run. Some agencies also simulate models submitted by

applicants. However, there are also open efforts to assess the need and means for a more thorough validation of machine learning models using techniques designed to interrogate machine learning models revealing the underlying reasoning or biases within the model.

Regulatory issues

Another area where agencies contribute to AI is guidance on regulatory issues. Some agencies report writing definition documents and providing guidance on frequently asked questions. They also provide regulatory advice on interpretation of legislation on matters involving the use of AI.

Examples

Please see the box below for some more detailed examples from the Paul Ehrlich Institute and the MPA:

Using AI in a regulatory setting

The Paul Ehrlich Institute is currently planning two projects involving the use of AI in a regulatory setting. This first project, RENUBIA, will attempt to peer inside the "black boxes" that are characteristic of machine learning models. The idea in the project is to develop and apply methods that can explain the underlying reasoning that machine learning models use to make their decisions. One method to be investigated will be interrogation of individual neurons in a convolutional neural network similar to the method developed by Google in the DeepDream project. Another such method will "attack" the machine learning models by subtly altering the input in such a way that the model now provides incorrect classifications.

The second project at the Paul Ehrlich Institute, KIMERBA, aims to facilitate the consistency of evaluations by providing assessors enhanced methods for interrogating historical applications and decisions. This project will apply natural language processing algorithms to extract and integrate information vital to regulatory decision making from the existing corpus of regulatory documents. This information will then be presented using a combination of visualisations and query interfaces allowing users to quickly find the information they are looking for. All information will be linked back to the original documents allowing assessors the ability to find the source.

The Swedish PhaVAI project was started in October 2021 in cooperation with Lund University and Stockholm University, with the overarching goal to improve quality and efficiency of the MPA processing of adverse event reports. In the currently ongoing phase, an ensemble of NLP triage models is being developed to identify serious events from the full-text event description, in order to prioritise these for manual review.

Challenges

Understanding and trust of AI

All participating EMRN agencies highlight issues regarding the scope and complexity of the AI field, and the use and trustworthiness of AI algorithms. Given the relatively recent spread of AI, its potential use cases remain unclear, along with ways it may supplement or supplant conventional methods.

The greatest challenge posed by AI to the network was the inability to understand the reasoning behind AI decisions. Many machine learning algorithms produce models that are “black box”; meaning that the description of the model is not easily interpretable by humans making it difficult to validate the quality of a model and to trust the predictions it makes. Furthermore, many studies show that the models can exhibit innate biases inherited from the data they were trained upon. The biases in machine learning models and their interpretability impacts both the in-house use of AI to support regulatory activities and the safety of medicinal products that use AI in their development or manufacture. When validating machine learning models, the sufficiency of using only performance measures needs to be investigated and, should they prove insufficient, the infrastructure and standards necessary to facilitate a more in-depth validation of machine learning models needs to be developed.

Another challenge is to understand the reliability of machine learning models when used outside the context in which they were developed. For example, models are often published in academic literature as prediction tools that are then integrated into workflows by third parties for the development and manufacture of a medicinal product. The generalisability of these models is not clear and methodology to assess their consistency and reproducibility would have to be established.

Communication

Communicating about AI presents several significant challenges. Even the term AI can have several different meanings depending on who is using it and the context it is used within. To facilitate the communication of AI related matters, several challenges will have to be overcome, including harmonising AI terminology, developing guidelines on how to best use and assess AI, and establishing standards to facilitate the transfer and use of AI data and models. In addition, the ability of an agencies to participate in AI related communication would depend heavily on whether they have personnel with sufficient training, guidance or expertise.

Infrastructure and capabilities

Currently, many agencies lack the infrastructure and capabilities to develop their AI capabilities. Although problems obtaining the necessary infrastructure for the evaluation and execution of machine learning models have been reported, obtaining and maintaining the necessary expertise is potentially more difficult. The field of AI is rapidly developing and personnel dedicated to keeping up-to-date with the latest developments would be necessary. To compound this problem, people experienced in AI are relatively rare and agencies will need to compete with each other and industry to obtain qualified experts. Finally, maintaining a high standard of expertise within an agencies may be difficult as departing personnel may leave specific areas of AI subsequently underserved.

Collaboration

Through collaboration within a CoE, many of the challenges posed by AI could be addressed. By sharing knowledge, expertise can be spread throughout the cluster. Harmonising terminology and establishing standards and guidelines will assist in sharing AI data and models, and provide a consistent and reliable approach for the use and validation of AI.

Knowledge sharing

Several possible opportunities are available that could foster a community of AI competency within the network. The most accessible approach would be to share knowledge, data and "soft" assets such as software and machine learning models. Sharing knowledge could come in several forms. Seminars on current research could keep the participating agencies up-to-date with the latest advancements, analyses and best practices in the field of AI as it applies to regulation. AI tutorials could provide the necessary understanding of how the different machine learning models are trained, tested and maintained enabling a better understanding of how to validate them. The code used in research and the development of new methods could be directly shared, allowing the efforts of one agencies to be reproduced in another. Finally, platforms designed to assist the regulatory process could also be shared allowing the capabilities of one agency to be replicated in another.

As many agencies lack the necessary infrastructure to train, execute and evaluate machine learning models, it has been proposed that computational infrastructure could be shared. This would come with its own set of challenges, but would potentially come with many benefits, such as the ability to provide evaluations that are consistent across agencies.

Multiple agencies proposed the development of a network of experts to tackle the issue of a lack of in-house experts. The agencies in the network could specialise on different AI aspects allowing the network, as a whole, to have a broad range of AI capabilities. Another proposed role of the cluster was to establish stable interdisciplinary teams of experts incorporating IT, data science, regulatory and medical-pharmaceutical expertise. Regular meetings and an online forum would augment knowledge transfer within the network and allow the network to collaboratively solve problems. Such a network could maintain a shared repository for AI related projects allowing each participating agencies to be aware of the other projects currently active within the network.

Harmonisation

A large area of responsibility perceived for a CoE in the area of AI is the harmonisation of terminology and assessments through the development of guidelines and standards. This work would not be confined to participants in the CoE, but would also address and be informed by planned BDSG work.

All aspects of an AI development pipeline can influence the quality of the resulting machine learning model such as its predictive power and generalisability. Given that such pipelines are highly complex, it is important that all aspects are properly assessed. Guidelines and requirements developed within the CoE would be valuable in assessing the different aspects of a pipeline from the data quality, provenance and integrity, to machine learning algorithm choice and training strategy. Guidelines could contain information about some of the common pitfalls of AI such proper procedures for choosing features, avoiding or detecting overfitting, optimising parameters and training a model. Such guidelines would provide a consistent approach to the assessment of AI use and result in harmonised guidance, regulatory and scientific advice, and interpretation of AI legislation.

Currently, there exist no standards for the exchange of AI related data or the application of AI to regulatory activities, inhibiting communication among parties interested in exchanging AI related information and slowing the adoption of AI capabilities. By collaboratively establishing a set of standards for data transfer, the CoE could augment data exchange among agencies and sponsors. Standards would have to cover a range of issues such as data quality, terminology, data formats, validation and software environments. When assessing AI supported products, standards would provide consistent evaluations across the different agencies. In addition, whether to assess sponsor trained models separately from published models needs to be decided.

Not all areas will be amenable to harmonisation. Even if expertise sharing and guidelines provide a consistent approach to assessment, there are still differences among the agencies that will prove more difficult to overcome, namely, country specific legislation. Countries such as Germany have quite strict data privacy laws that will impede the transfer of data especially where patient data is concerned.

Appendix A - Abbreviations

AEMPS - Agencia Española de Medicamentos y Productos Sanitarios

ADaM - Analysis Data Model

AI - Artificial Intelligence

BDSG - Big Data Steering Group

BfarM - Bundesinstitut für Arzneimittel und Medizinprodukte (The Federal Institute for Drugs and Medical Devices)

CDISC - Clinical Data Interchange Standards Consortium

CHMMP -

CoE - Cluster of Excellence

DAC - Danish Medicines Agency's Data Analytics Centre

DKMA - Danish Medicines Agency

EHR - Electronic Health Record

EMA - European Medicines Agency

EU-NTC - European Union Network Training Center

GDPR - General Data Protection Regulation

HMA - Heads of Medicines Agency

MEB - Medicines Evaluation Board, Netherlands

ML - Machine Learning

NCA - National Competent Agency

PEI - Paul Ehrlich Institute

RWD - Real World Data

RWE - Real World Evidence

SDTM - Study Data Tabulation Model