HMA-EMA Joint Big Data Taskforce Phase II report:

'Evolving Data-Driven Regulation'

[1] Corr. The links in the Annex IV and Annex VII have been updated.

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

Table of contents

1. Executive summary

As healthcare data and technology evolve, then so must medicines regulation. This report "Evolving Data-Driven Regulation" represents Phase II of the HMA-EMA joint Big Data Task Force (BDTF). It prioritises recommendations from Phase I of the work and suggests ways forward for the European regulatory network and stakeholders to realise the potential of Big Data in terms of public health and innovation, through evolution of our approach to using data to generate evidence. The report aims to inform strategic decision-making and planning by the HMA and EMA and to input to the EU Network Strategy to 2025. The report will support regulators and stakeholders seizing the opportunity for data-driven, evidence-based, robust decision-making that will underpin the development, authorisation and on-market safety and effectiveness monitoring of medicines in a rapidly evolving data and analytics landscape.

The increasing volume and complexity of data now being captured across multiple settings and devices coupled to rapidly developing technology offers the opportunity to deliver a better characterisation of diseases, treatments and the performance of medicinal products in individual healthcare systems. Such data sources, commonly labelled as Big Data, are generally large, accumulating rapidly, incorporate multiple data types and forms, and are of varying value and quality. Big Data includes real world data such as electronic health records, registry data and claims data, pooled clinical trials data, datasets from spontaneously reported suspected adverse drug reaction reports, and genomics, proteomics and metabolomics datasets. Big Data can complement clinical trials and offers major opportunities to improve the evidence upon which we take decisions on medicines. Understanding the quality and representativeness of Big Data will allow regulators to select the optimal data set to study an important question impacting the benefit-risk balance of a medicine. Establishing the IT capability and capacity to receive, manage and analyse Big Data will enable the Network to discover insights on the safety, efficacy and use of medicines and explore the validity of claims made by the industry. Building expertise to advise, interpret and analyse Big Data will ensure the EU Network can both meet the challenges of product dossiers which include such data and, moreover, realise the public health and innovation benefits of Big Data. Guiding Big Data's use by industry, understanding the Big Data evidence submitted, and conducting Big Data analyses will get medicines to patients more quickly and optimise their use on the market.

The BDTF worked in 2017 and 2018 on its Phase I report which reviewed the landscape of Big Data and identified opportunities for improvement in the operation of medicines regulation. The Phase I report was published in early 2019 and was used to stimulate feedback from stakeholders. In parallel with the consultation, the BDTF started Phase II of its work with a particular focus on prioritising the recommendations from Phase I and making practical suggestions on how to execute the recommendations and how the European regulatory network could collaborate with stakeholders to realise the potential of Big Data.

The recommendations of Phase II of the BDTF are organised into those to be implemented through collaboration with stakeholders, those for action by the European medicines regulatory network, and those for action by specific committees or working parties of the EMA. From the large number of recommendations identified in Phase 1 of the BDTF and elaborated in Chapter 5 of the present report, Phase II has distilled 10 priority recommendations which are fully compatible with the current EU legal framework for the regulation of medicinal products.

i. **Deliver a sustainable platform to access and analyse healthcare data from across the EU (D**ata **A**nalysis and **R**eal **W**orld **I**nterrogation **N**etwork -**DARWIN).** Build the business case with stakeholders and secure funding to establish and maintain a secure EU data platform that

supports better decision-making on medicines by informing those decisions with robust evidence from healthcare.

ii. **Establish an EU framework for data quality and representativeness**. Develop guidelines, a strengthened process for data qualification through Scientific Advice, and promote across Member States the uptake of electronic health records, registries, genomics data, and secure data availability.

iii. **Enable data discoverability**. Identify key meta-data for regulatory decision-making on the choice of data source, strengthen the current ENCePP resources database to signpost to the most appropriate data, and promote the use of the FAIR principles (Findable, Accessible, Interoperable and Reusable).

iv. **Develop EU network skills in Big Data.** Develop a Big Data training curriculum and strategy based on a skills analysis across the network, collaborate with external experts including academia, and target recruitment of data scientists, omics specialists, biostatisticians, epidemiologists, and experts in advanced analytics and AI.

v. **Strengthen EU network processes for Big Data submissions.** Launch a 'Big Data learnings initiative' where submissions that include Big Data are tracked and outcomes reviewed, with learnings fed into reflection papers and guidelines. Enhance the existing EU PAS register to increase transparency on study methods.

vi. **Build EU network capability to analyse Big Data**. Build computing capacity to receive, store, manage and analyse large data sets including patient level data (PLD), establish a network of analytics centres linked to regulatory agencies, and strengthen the network's ability to validate AI algorithms.

vii. **Modernise the delivery of expert advice**. Build on the existing working party structure to establish a Methodologies Working Party that encompasses biostatistics, modelling and simulation, extrapolation, pharmacokinetics, real-world data, epidemiology and advanced analytics, and establish an Omics Working Party that builds on and reinforces the existing pharmacogenomics group.

viii. **Ensure data are managed and analysed within a secure and ethical governance framework.** Engage with initiatives on the implementation of EU data protection regulations to deliver data protection by design, engage with patients and healthcare professionals on data governance, and establish an Ethics Advisory Committee.

ix. **Collaborate with international initiatives on Big Data.** Support the development of guidelines at international multilateral fora, a data standardisation strategy delivered through standards bodies, and bilateral collaboration and sharing of best practice with international partners.

x. **Create an EU Big Data 'stakeholder implementation forum'**. Dialogue actively with key EU stakeholders, including patients, healthcare professionals, industry, HTA bodies, payers, device regulators and technology companies. Establish key communication points in each agency and build a resource of key messages and communication materials on regulation and Big Data.

Implementation should be overseen and success factors defined and measured by an HMA-EMA Steering Group on Big Data.

Putting the recommendations into practice will involve: training and developing our staff and targeted recruitment of new staff; delivery of demonstration pilots; establishing analytics and regulatory science centres of excellence; developing guidance; strengthening existing regulatory tools such as qualification advice; investing in fit-for-purpose targeted information technology, and; delivering a bold Big Data initiative (DARWIN) to establish a framework for accessing and analysing EU healthcare data with an initial focus on real-world data.

In moving forward, success factors will include building on the strengths of the current system, working collaboratively within the EU regulatory network and with EU and international stakeholders, providing clear requirements for the regulated industry, and targeting the network's efforts to where the maximum health and innovation benefits can be delivered.

Big Data is not necessarily the solution to all the challenges faced by regulators in reaching appropriate decisions. While randomised, double-blind, controlled clinical trials will remain the reference standard for most regulatory use cases, the complementary evidence that new Big Data sources generate may facilitate, inform and improve our decisions. It is clear that the data landscape is evolving and that the regulatory system needs to evolve as well. In this way we can realise opportunities for public health and innovation through better evidence for decisions on the development, authorisation and on-market safety and effectiveness monitoring of medicines. If we work now, smartly and collaboratively, and embrace change we can evolve to deliver better regulation for patients and establish the EU medicines regulatory network as a reference for data-driven decision-making.

2. Introduction

Science and technology are developing at an unparalleled rate. The convergence of new treatments, diagnostics, wearables, sensors and connectivity is generating enormous amounts of data. Data from clinical trials are becoming available for re-analysis and cross-linking to other data sources[2,3] and, omics-driven methods are used in precision medicine and in innovative, individualised therapeutic approaches. Computational power and approaches based on bioinformatics tools and algorithms, machine learning (ML) or artificial intelligence (AI) are gaining access to the health care systems. Likewise, our possibilities for evidence generation are expanding. For example, while the introduction of electronic patient records, which aim at recording and making accessible a patient's journey, began in some countries nearly 20 years ago, it is only recent advances in information technology that have created the infrastructure that allows these data to be used by enabling data to be securely aggregated, stored, processed and transmitted. The combination of these drivers is resulting in a sea change in data availability, offering new opportunities for evidence generation. Layered on top of this, the scientific environment in which medicines are developed and delivered is changing fast and the pace of change is likely to accelerate further and will increasingly impact on the way our health care is delivered. In the future, medical care and diagnostics are likely to rely more and more on data-driven technologies such as AI/ML for disease diagnosis, wearable devices and sensors to assess basic physiological parameters, patient activities and multiple biomarkers to monitor disease as well as its progression and response to treatment. While validation of approaches will be required, it is likely that in an era of precision medicine, a diagnosis and prescribed treatment may depend not only on your genome but also on your epigenome, proteome, microbiome and metabolome as well as your behaviourome (i.e. factors related to physical activity, nutrition, mental health etc.) with a view to identifying the right treatment, at the optimal point in the disease at an individual patient level.

> *Big data may only tell you what works rather than why*
>
> *Public consultation comment*

Evidence generation needs to keep pace. Trials are typically multi-centre and multi-national to meet the need to include cohorts from countries where the drug will be marketed and reflect ethnic and cultural diversities (as well as ensuring appropriate patient numbers are included). The emphasis on clinical outcomes as opposed to surrogate markers adds to the duration and size of trials (despite selection of patient populations) and contributes to the fact that in the decade from 2002 to 2012 the number of endpoints per trial has nearly doubled, and the average number of procedures that a trial participant underwent increased by 58% [1]. Any differing requirements of multiple independent regulatory agencies and subsequently of Health Technology Assessment bodies and payers drives further complexity and cost. Despite such investments and increases in complexity, we continue to see late stage failures in drug development - nowhere more apparent than in the Alzheimer's field [2].

So, in the face of these changes in data generation and scientific innovation, is our current drug development model – and regulatory paradigm – sustainable, and if not, how must we adapt? It could be argued that the fundamental regulatory model has remained largely the same for decades. Of course, policies and processes have changed significantly: the launch of the European authorisation system; new approaches to allow accelerated access in the face of unmet medical need; new pharmacovigilance legislation to provide regulators with more powerful tools to demand studies on safety and efficacy post-authorisation; new legislation for advanced therapy medicinal products[4], new incentives to promote medicine development in rare disease or paediatrics; substantially increased

[2] https://yoda.yale.edu/yoda-project-metrics
[3] https://www.clinicalstudydatarequest.com/Metrics.aspx
[4] https://www.ema.europa.eu/en/human-regulatory/overview/advanced-therapies/legal-framework-advanced-therapies

transparency of the data submitted in support of a medicines efficacy and safety, and processes to increase engagement with all actors in the healthcare market and most recently new legislation on medical devices and in vitro diagnostic devices[5]. Europe has also been a pioneer in establishing processes to engage and involve patients in regulatory decisions, an approach that was initiated partly through lessons learned in establishing early access to human immunodeficiency virus (HIV) treatments[6]. Nevertheless today, in Europe, medicines are still approved following the assessment of summary data submitted by the marketing-authorisation applicant and independent regulatory assessment of the data at the patient level is not normally performed. The randomised controlled trial is still considered the best available standard for the assessment of efficacy and in the context of a regulatory application, the generation of that data follows strict guidelines, is verified at source, and the evidence generated on the basis of pre-specified analyses agreed prior to the start of data collection. Post-authorisation safety is still largely assessed through the submission of adverse drug reaction (ADR) reports from healthcare professionals, and although the new pharmacovigilance legislation introduced new tools to monitor the benefits and risks of products, ADR reports remain the main source of new drug safety signals. Moreover, the full benefit-risk balance is principally assessed once at the time of marketing authorisation, with ongoing assessment focussing mainly on the risk (safety) side of the equation. Maybe as a consequence of this, changes to a product authorisation are still usually applicant driven unless a specific safety issue is identified.

Are there opportunities to improve or refine our decision-making? Is more data the answer? Can we better use and analyse existing data? While concerns have been voiced for years around the generalisability of the clinical-trial data to normal clinical practice, given the robust selection criteria and selective trial environments, to date there has been no better alternative to replace the evidence generated by trials. However, the sea change in not only data availability, but in its variety, depth, detail, quality and source have changed the landscape and is creating pressures on regulators to have a clear position on, not only when and where these data may be acceptable for their decision-making but to provide clear metrics for applicants to understand the reasons on which these decisions are based. There are uncertainties of quality and of bias and the report from Phase I of the BDTF set out a range of activities at the wider community level, which aim to improve the quality and trustworthiness of the data and subsequent evidence. However, novel data generation approaches should not be rejected on the basis of subjective concerns about the data, principally that the quality is not sufficient or that there are unknown biases in order to avoid tackling the issue. Such concerns may have substance but need to be based on fact rather than perceptions. A pertinent quote from the BDTF's Big Data solutions meeting held in 2018 was 'Defensive organisations rarely try new things[7]'. While we must not disrupt a functional regulatory model which is delivering robust and proven secure decision-making, equally we must not be afraid of change which, if managed and implemented appropriately will ensure that the EU regulatory system is ready for the challenges of the future".

So the challenge is how does one design a regulatory model that can capitalise on the promise of additional evidence from novel datasets of unknown quality and provenance where pre-specified statistical analysis may not be possible, and yet still reach a robust, assured position of the benefit-risk of a medicine? There are many questions: How do we efficiently integrate data analysis into our assessment processes to improve and refine our decision-making? When and how should the regulator re-analyse data to verify a key finding? How do we incorporate complex, personalised, new information in the intended target population arising from these data in real time, which potentially affects the benefit-risk of these medicines into our product information? In the era of precision medicine, how do

[5] https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices
[6] J Ambul Care Manage. 2010 Jul-Sep;33(3):190-7. The Patients' and Consumers' Working Party at the European Medicines Agency: A Model of Interaction Between Patients, Consumers, and Medicines Regulatory Authorities. Isabelle Moulon; Nikos Dedes
[7] James Kugler – Merck

we understand benefit, and manage risk at the individual level rather than the population level? What will be the threshold of evidence arising from Big Data analyses to determine regulatory action? How do we upskill the regulatory workforce such that there is the expertise to critically assess the evidence arising from these data? And in the face of these changes, how do we design a model that extracts value out of Big Data to ensure the implemented processes and mechanisms are sustainable for the regulatory system. Finally, in all our work, we need to be sure that the evidence is reliable and our decisions are robust as these are essential to building trust of patients and healthcare professionals in the regulatory system and, ultimately, in the medicines on the market. These challenges are encapsulated in the problem statement which has guided the Big Data BDTF in building its recommendations.

2.1.1. Problem statement

Advances in information technology are driving digitisation of large volumes of often unstructured research and clinical data, commonly termed Big Data. While the capture and analysis of these data offer possibilities to derive novel insights, the acceptability of such insights as evidence for regulatory decision-making needs to be clarified.

Frequently, pre-specified, standardised analyses of Big Data are not possible and changes in approach and additional assumptions are required. In addition, additional re-analyses of Big Data sets may be needed to validate results and ensure confidence in the derived conclusions. Currently the EU regulatory network has limited capacity and capability to access and analyse large and unstructured data sets, and needs to be strengthened to guide the use of emerging technologies and critically interpret analyses based on Big Data or novel analytical approaches.

3. Phase I of HMA-EMA Joint Big Data Taskforce

In the first phase of its work, the BDTF delivered on its mandate in order to:

- map relevant sources of Big Data and define the main format, in which they can be expected to exist and through a regulatory lens describe the current landscape, the future state and challenges;
- identify areas of usability and applicability of emerging data sources;
- perform a gap analysis to determine the current state of expertise across the European regulatory network, future needs and challenges;
- generate a list of recommendations and a Big Data Roadmap.

Six subgroups were initially formed to describe from a regulatory perspective the characteristics and potential areas of usability of genomic data, bioanalytical omics (predominantly focussed on proteomics), clinical-trial data, observational data, spontaneous ADR reports, m-health and social media data. At a later date, a data analytics subgroup was formed as all other subgroups identified Big Data analytics as an important area of focus. In addition, the BDTF undertook surveys of (i) the European regulatory network to assess the available expertise and competences for analysis and interpretation of Big Data and (ii) an e-survey of pharmaceutical companies seeking to understand the current experience, key challenges, applicability and added value of Big Data over the life cycle of a product. As the main deliverable of Phase I, the BDTF generated a summary report [8] and 7 subgroup reports which together generated 47 core recommendations and 138 supporting reinforcing actions.

[8] https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf

The specific recommendations were presented in Annex III of the summary report[9]. The summary report organised these recommendations around the principles of data standardisation, data quality, data linkage and data analytics in addition to horizontal cross-cutting recommendations addressing medical devices and in vitro diagnostics and training skills and communications. Notably the report set out 'what' needed to be addressed but highlighted that 'the how' and 'the when' required further work.

3.1. Stakeholder responses to the summary report

The summary report and table of recommendations were published externally on HMA and EMA websites in February 2019 with a consultation period of 2 months in order to gather comments from external stakeholders. Thirty-eight responses were received during the consultation period.

A synopsis of the key points is provided at Annex I.

4. Phase II of HMA-EMA Joint Big Data Taskforce

4.1. Taskforce structure

The seven subgroup reports from Phase I generated 47 core recommendations and 138 supporting reinforcing actions[10]. As a result, it was clear that a prioritisation and focussing of activities was required to move from more high-level recommendations to practical and concrete actions. This was the mandate of Phase II of the BDTF, coupled with an estimate of the required resources for implementation.

Many of the recommendations from Phase I had common themes and it was clear that if the top down, data-focused approach was continued in Phase II there was a risk of significant duplication of effort. Consequently, all recommendations and associated reinforcing actions were stratified according to themes which led to the creation of six new horizontal cross-cutting subgroups to address the following areas:

- Data Sharing/accessibility
- Data standards, quality and infrastructure
- Data Analytics
- Devices and In vitro Diagnostics
- Regulatory Acceptability
- Research Initiatives
- Policy, training and Communications.

The BDTF is currently co-chaired by Peter Arlett (EMA) and Nikolai Brun (HMA, Denmark). Alison Cave (EMA) co-chaired until September 2019. The full membership for Phase II of the BDTF can be found at Annex II.

4.2. Methodology

It is clear that the ownership and means to deliver many of the recommendations is not solely within the mandate of the European regulatory network. In order to progress the work and prioritise the

[9] https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf
[10] https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf

recommendations the BDTF has stratified the original recommendations into three main areas based on how they should be delivered.

Collaborating:

> The recommendation requires multi-stakeholder alignment and co-ordinated action and cannot be delivered by the European regulatory network alone. However, the network has significant opportunity to collaborate and influence initiatives to ensure outcomes that benefit public health.

European regulatory network level:

> The recommendation requires agreement at the level of the network and subsequent consolidated action.

***NCAs, EMA Committees or working party*:**

> The recommendation can be delivered by a single NCA, an EMA Committee or working party. That is, the recommendation does not require consolidated action across the entire European regulatory network.

Recommendations were allocated across the six new subgroups, and members were asked to complete an assessment fiche for each recommendation (see Annex III for fiche structure). A list of fiches which drive key recommendations is provided at Annex IV, and accompanying documents are available. It should be noted that some fiches cover more than one recommendation.

5. Recommendations

Recommendations from Phase I and II of the BDTF should be viewed as a whole. Phase I recommendations relate more to the wider landscape and delivery will generally require consolidated action of multiple stakeholders and substantial resources. Recommendations from Phase II build on that foundation and focus on preparing our regulatory model for the new data environment.

The problem statement articulates a number of key limitations likely to develop with the current regulatory paradigm as large volumes of research and clinical data become available. More specifically, it focuses on the limited capacity and capability currently within the European regulatory network to access and analyse large, heterogeneous and unstructured data sets. In order to address this, the BDTF has prepared the following vision statement which articulates a number of key elements required to build a regulatory network that is future proofed for the evolving scientific and regulatory landscape. In considering recommendations to deliver the vision, the BDTF has been informed by reviews in the literature, the draft EMA Regulatory Science Strategy to 2025 and the stakeholder comments submitted during its public consultation.

5.1. Vision Statement

The vision of the BDTF is of "a **strengthened regulatory system** that can efficiently **integrate data analysis** into its assessment processes to improve decision-making. This will be supported by **knowledge of data sources, their quality and their relevance for the European population**, **continual optimisation of data quality and analytical approaches** and promotion of a **secure and ethical data sharing culture**. **Training and external collaborations** will be key in order to build expertise.

Knowing when and how to rely in novel technologies, and the evidence generated from Big Data, will benefit public health by accelerating medicines development, improving treatment outcomes and facilitating earlier patient access to new treatments."

The following set of recommendations are specifically aimed to strengthen our regulatory paradigm to enable it to understand and use the available data and equally respond to uncertainties raised by others with the data. This will require agile and flexible processes to deliver appropriate guidance and consistent decisions to ensure one set of uncertainties are not replaced with another. The recommendations are compatible with the existing legal framework although some could be more impactful with an explicit legal basis.

Within a detailed table provided in Annex V recommendations are organised into the six parallel areas aligned with the vision statement, and further stratified according to by whom the action may be delivered. A short synopsis of each recommendation is provided below with more detail of key recommendations available in working BDTF documents ("fiches").

5.2. Establish a framework which describes data quality

Establish a certification process for data sources '

'Data quality is not a static construct and is context, disease and question dependent and dependent on the healthcare system. Assessments need to be constant and documented every time the data is refreshed'

Public consultation comments

The Big Data vision requires the use of data not originally intended for regulatory decision-making and understanding quality is challenged by a lack of standardisation, sometimes limited precision and robustness of measurements (e.g. proteomics data), missing data, variability in content and measurement processes, unknown quality and constantly changing datasets. One possible exception to this is the well standardised adverse drug reporting datasets. As an example a recent analysis revealed that the number of European databases that meet minimum regulatory requirements for content across a broad range of regulatory use cases and which are readily accessible is disappointingly low [3]. See figure 1 for illustration of the data landscape for real-world data.

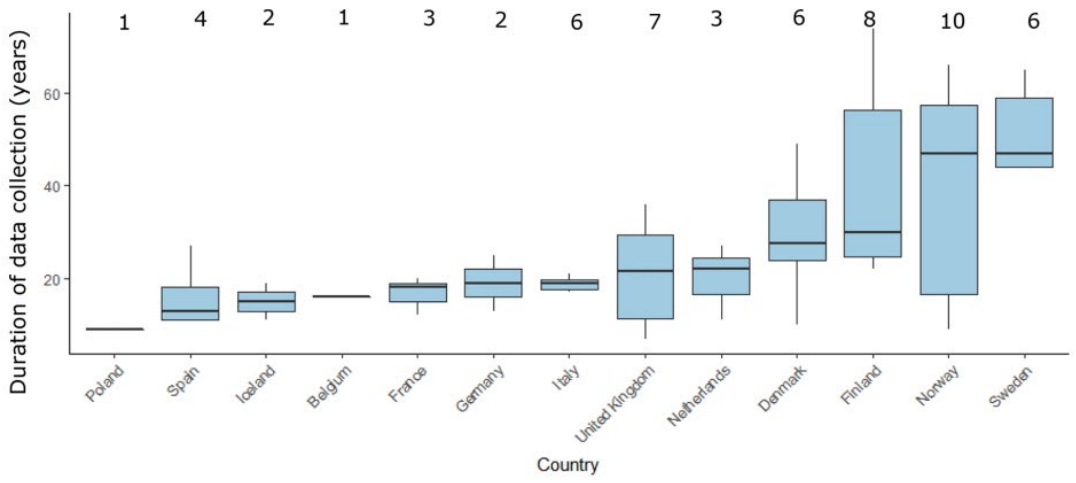

Figure 1: European data sources and duration of data collection. Box plots indicate the median (horizontal black line) data collection time by country while the margins of the box plot represent the IQR, the vertical lines indicate the minimum and maximum values. The numbers of databases per country are provided above the box plots [4].

In order to include novel data sources as evidence sources for regulatory decision-making, it is critical to understand how much the regulators can rely on the data. Thus, a capability to characterise the quality of data is a strategic objective for regulators. While pre-defining quality is challenging as need is often driven by the question, it is possible to define some generalised elements for which quality could be defined.

The establishment of renewable certification processes will be key for ensuring trust and aiding interpretability of studies generated using such data; expansion of qualification advice is an obvious regulatory route to achieve this. Such certifications should additionally provide knowledge on which actions are required to improve the reliability of the data. It will also be important to incorporate data quality needs from Health Technology bodies and payers to create a framework with the widest possible utility. Consideration should be given, as to where it is possible to build on the work of international regulators in order to reach an aligned position; for example the FDA's April 2018 final guidance "Use of Public Human Genetic Variant Databases to Support Clinical Validity for Genetic and Genomic-Based In Vitro Diagnostics" provides a mechanism for test developers to leverage publicly accessible databases of human genetic variants to support the FDA's regulatory review of genetic and genomic tests. FDA recognition of a database indicates that the FDA believes the data and assertions contained in the database can be considered valid scientific evidence and will allow sponsors to use the assertions within FDA recognised databases to support the clinical validity of their tests.

> *'EMA and other regulatory agencies need to work in a timely manner with pre-competitive consortia and providers of RWD to co-produce data and data standards that are fit for purpose'*
>
> *Public consultation comment*

Recommendations

Establish a data quality framework (DQF) for regulatory use of big data sources with associated data quality metrics

Expansion of qualification advice process to establish renewable certification of datasets as well as Big Data methods and strategies

Establish criteria for reliability of device based diagnostic and other in vitro diagnostics

Proactive external communication to promote adoption of Data Quality Framework

Promote use of ISO-IDMP standard

Fiche #1 & 2

What this means for stakeholders:

A data quality framework will support the trust of patients and healthcare professionals in the decisions reached by regulators when Big Data underpins those decisions. It will aid the choice of data source selected for a study (including those by industry) and it will inform the assessment of the study results and the benefit-risk dossier by regulators.

5.3. Define a strategy to assess the representativeness of relevant data sets

It is anticipated that science will offer the possibility to better stratify diseases via for example detailed molecular profiling and diagnostic imaging, which is likely to result in evidence presented at authorisation in smaller and more defined patient populations. Increasingly this may involve data

captured by mobile devices which provides individualised geographical and lifestyle information. Understanding the applicability of this evidence to a wider population will be key and will require access to multiple complementary sources of evidence to proactively track benefit-risk over the entire life cycle of a product. Similarly, at a national level healthcare data are heterogeneous as differences in healthcare systems, national guidelines, and clinical practice have driven different content and historically different healthcare systems have used different terminologies and structures. For other stakeholders such as HTA bodies and payers, access to data representative of the relevant Member States may be required as the generalisability of evidence derived in different MS for economic decisions may be challenging. To meet these disparate needs access to timely data of sufficient quality and scope and representative of the entire European patient population will be critical.

Recommendations

Promote the development of data sources that might be used for analytical purposes in Member States where currently there are none

Develop guidelines for acceptability of evidence derived from different populations and jurisdictions for European submissions

Support and promote initiatives to access and link across care settings

Fiche #3

Representativeness does not only relate to the country of origin but also to the healthcare sector in which the care is delivered. This is illustrated by a recent pilot performed by EMA in the context of signals reviewed by PRAC where only 53.4% of centralised authorised products considered over a 3-month period had an adequate exposure for study in at least one of the primary healthcare datasets interrogated. This percentage decreased significantly if frequency is required in at least 2 or 3 of the possible databases currently available to EMA in house. Thus, access to data of an adequate quality from secondary and tertiary care is an urgent need. Lastly, it is envisaged that increasingly the European regulatory network will be asked to consider data derived from non-EU datasets in regulatory submissions and the appropriateness of this data will need to be determined.

5.4. Improve Data discoverability

Identification of appropriate data sources is becoming an increasing need in regulatory decision-making, for instance in the context of long-term follow up of innovative medicines and other post-authorisation obligations for products authorised by a conditional authorisation. In addition, data needs are becoming ever more complex. For example, one pillar of the long-term follow-up of the new tumour agnostic cancer therapies will be the collection of mutation specific clinical outcome data across multiple histological subtypes of cancers and hence likely multiple disease registries. Finding suitable data sources to deliver data of sufficient depth and detail in several European Member States will be very challenging and resource intensive. Failure to identify suitable data sources leads to the establishment of new sources to meet post-marketing obligations with the resultant duplication of effort, loss of resources and further fragmentation of the data landscape.

However, discovering data is a challenge; data sets are often siloed by country, language, region, hospital and even department and almost always captured in a disease specific context. Moreover data are often inaccessible, annotated inconsistently and recorded using multiple different terminologies which are often further modified at the local level. Data discoverability is thus a huge challenge for the entire scientific community. In recognition of this, the FAIR principles were developed in 2016 [5] to make data Findable, Accessible, Interoperable and Reusable. The principles emphasise machine findability as a fundamental principle. For Big Data approaches it is essential that automated data processing algorithms efficiently identify appropriate data sets, based on the information provided by

the meta-data of a respective data set. The meta-data have to have a format that is 'readable' by the used data processing algorithm.

FAIR principles suggest that metadata should be generous and extensive, and should include information about the context, quality, and condition, or characteristics of the data and should not pre-suppose a purpose or user for the data. The principles of FAIR have gained much traction and momentum including recent endorsement by global organisations such as G7[11]; in addition, the Innovative Medicines Initiative has recently funded FAIRPLUS which aims to develop tools and guidelines for making life science FAIR[12]. It is key that the regulatory network engages with this initiative and builds consensus around data characteristics, be it content, context or quality, relevant to regulatory needs. This long-term objective will improve data discoverability in the future but is unlikely to meet immediate needs.

Recommendations

Harness FAIR principles and processes as a mechanism to sustainably increase data discoverability

Include key data elements specific for regulatory needs in FAIR Data Points

Expand the scope and utility of the ENCePP database to improve discoverability of healthcare data sources

Increase horizon scanning efforts to identify relevant data sources

Fiche #4

To meet short-term needs (1- 5 years), expansion of the scope and utility of the ENCePP resource database could provide detailed information on source, spectrum and quality of datasets. While currently focussed on RWD, the scope could embrace other healthcare data types. It is recognised that other inventories are available for example EMIF (European Medical Information Framework), but the sustainability of such inventories are not assured beyond the lifetime of the specific projects. Moreover, a regulatory hosted inventory provides a trusted and independent environment which is more likely to gain the trust of multiple stakeholders.

For other Big Datasets, engagement with relevant European and Global initiatives such as Genomic Alliance for Genomic Health (GA4GH) and Elixir will raise regulatory awareness around dataset availability and relevance for decision-making. For example, GA4GH have a number of driver projects which may generate useful data to aid regulatory decision-making e.g. BRCA Challenge[13], Autism Sharing Initiative[14], ICGC-ARGO[15] but of which there is very limited awareness.

Inclusion of a description of datasets in the EU Network Training Centre (EU NTC) as a central resource would provide a useful mechanism to promote knowledge exchange across the regulatory network.

What this means for stakeholders:

Increasing data discoverability will help signpost industry and regulators to the best data source to address a particular regulatory use case whatever the regulatory procedure (from preauthorisation drug development to on market performance monitoring. Increased discoverability will ultimately improve the evidence available to reach benefit-risk decisions and facilitate getting better medicines to patients.

[11] http://www.g8.utoronto.ca/science/2017-annex4-open-science.html
[12] https://fairplus-project.eu/
[13] https://brcaexchange.org/
[14] https://www.autismsharinginitiative.org/
[15] https://icgcargo.org/

5.5. Further strengthen the robustness of decision-making

'FDA guides data providers on what data is needed for informing decisions; EMA needs to consider adopting the same approach'

'The phase I report lacks emphasis on the increased risk of bias in many big data approaches, there needs to be more awareness of both regulators and public, on this underlying, hidden but deleterious risk'

Public consultation comments

The creation of robust data quality frameworks across multiple datasources will significantly support decision-making but the evidential requirements are not equivalent for all regulatory decisions or at all stages of the product life cycle. Acceptability is influenced by the context; the question being asked, the level of risk associated with each decision and other considerations such as the ability to capture other data, availability of other treatments and unmet medical need. Moreover, acceptability of evidence is not only dependent on data quality but also on the methodological processes used to generate the evidence and on the measures implemented to control for bias and confounding; this is especially true for evidence generation in the absence of randomisation, blinding or use of an appropriate comparator. Defining acceptability is further challenged when considering the threshold for evidence, which would be required to initiate regulatory action in response to putative causal associations arising from Big Data analyses. Guidance is thus needed to inform on regulatory expectations around data quality, other relevant data attributes and methodologies across the range of regulatory decisions in the context of the risk associated with each decision, which should, where possible, seek alignment with other regulatory authorities.

Recommendations

A strategic initiative to gather learnings on the utility of RWE and Big Data in drug development which includes a dedicated RWE pilot programme from which lessons should be publicly shared

These learnings support an iterative framework which defines the evidential requirements for acceptability of new sources of data across the full range of regulatory decisions

Develop guidelines on study design and reporting and require the public posting of protocols, amendments and results for studies for regulatory submissions

Fiche #5 & 6

Guidance will be significantly informed by a strategic initiative to gather learnings on the utility of Big Data in drug development. As an example, given the maturity of the field, a dedicated RWE pilot programme is recommended where stakeholders are invited to submit demonstration projects across a range of product types and diseases for detailed regulatory advice from which the lessons learned should be publicly shared. Proposals using a range of data sources, especially a combination of different types of Big Data, would be especially encouraged. This initiative should assimilate building blocks across the commonly available regulatory tools (e.g., guidance, pilots, capability building, and stakeholder engagements) to develop guidance on what factors should be considered and addressed in a regulatory submission.

'More pilot studies and collaborations to inform thinking'

Public consultation comment

As 'omics become increasingly important in prescribing decisions, consideration must be given to the level of clinical evidence required to extend an indication to a wider population based on the validation of new biomarkers. Equally, as Big Data analyses accelerate, more pharmacogenomics markers will be determined and guidelines are required as to the level of clinical evidence required for incorporation

into the product information. Guidelines on evidentiary standards should include a consideration of such challenges.

What this means for stakeholders:

By analysing the submission of Big Data to regulators we can inform the development of guidance that supports better and more insightful use of Big Data in the future. An iterative framework will support training and capacity building for industry and regulators and help to support better decision-making on products for patients.

5.6. Ensure the efficient and targeted integration of Big Data analysis into the decision-making process

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 from a large variety of heterogeneous sources requires new technologies and architectures to analyse and generate the anticipated value.

> *'Guidance on the acceptability on the use of data from different global regions.'*
>
> *Public consultation comment*

The European regulatory system does not routinely require submission of raw data (patient level data - PLD) from clinical trials. For regulatory applications it is currently required to submit confirmatory (a priori) analysis of clinical data including pre-specifying documents, such as clinical study protocols and statistical analysis plans. The assessment of the results and the pre-specifying documentation in combination with inspections has generally been considered sufficient, to ensure robustness of evidence. Thus, in the past there has been no routine secondary analysis established by the competent authorities, to scrutinise the underlying PLD and the data analysis.

However, if the data is such that pre-specified, standardised analyses are not possible, re-analysis of the data by the regulators becomes necessary, in order to ensure the validity of results and the appropriateness for regulatory decisions. Other international regulatory agencies, including the US FDA and Japanese PMDA, already receive PLD as part of regulatory submissions and use it to support their assessment of the marketing-authorisation application dossier. Both agencies mandate CDISC standards for datasets and associated metadata for marketing-authorisation applications. If we are to move towards a system where we can efficiently integrate data analysis into our decision-making, we need a change in approach and new processes need to be piloted.

Recommendations

Modernise IT infrastructure to enable regulators to be able to securely collect, store, manage, explore, link and analyse Big Data sources in an efficient, secure, adaptive and scalable manner

Formation of a cross committee taskforce to examine the practical aspects of PLD analysis with an initial focus on clinical-trial data

Enriched scientific advice related to big data applications where there may be significant uncertainties which require PLD analysis to resolve

New cross committee methodology working party building on existing groups and enriched with RWE and advanced analytics

Strengthened 'omics working party building on existing pharmacogenomics working party

Develop strong and systematic links between regulatory agencies responsible for medicinal devices/products and notified bodies

Fiche #7, 8, 9, 10

Provision of sufficient computational capacity and relevant expert skills such as data managers, data scientists and statisticians to analyse Big Data sets in a timely fashion is fundamental to delivering the vision, and must be supported by proper management processes, investments and data governance.

It is noted that in the Information Management Strategy 2019-2021, the network already foresees the need to manage and analyse potential Big Data sources and hence this recommendation complements and supports this strategy.

The question of whether individual PLD[16] should be assessed as part of the authorisation procedure was considered by EMA Management Board in December 2014. At the time, the Management Board considered a deeper reflection was required to clarify of the objectives and develop criteria which would trigger a patient level analysis and an examination of resource implications was requested.

Due to the rapid pace of digitalisation, the data landscape has changed significantly since that request, and even more change is anticipated in the coming years. Hence, as data availability increases and data from more sources are integrated into applications, it is the view of the BDTF that our current approach will become increasingly limited and will potentially impact on the robustness of our assessments.

The regulatory environment needs to be prepared to validate and ensure reliability, quality and regulatory compliance of the data

Public consultation comment

Thus, the BDTF fully supports the potential added value and benefits of PLD analysis for the evaluation of benefit-risk of human medicines, upon request from EMA Scientific Committees, and within the boundaries set by the current applicable legislation. A recent position paper from the Biostatistics Working Party (Annex VI) sets out two key recommendations adopted by the BDTF namely a cross committee group to examine practical aspects of PLD analysis and a Proof of Concept pilot to inform the estimation of human resourcing and technological needs. The proof of concept pilot should also examine the scenarios in which PLD assessments might add value. More detail around this recommendation is provided within the position paper. Increasingly this approach will be extended to other contexts for example validation of machine learning algorithms especially in the context of composite digital endpoints.

PLD analysis has already been seen in qualification advice[17] for a new prognostic biomarker for autosomal dominant polycystic kidney disease where the raw data was requested in order to allow a better understanding of the competence of the database and the model. Such instances are expected to increase and it is anticipated that a consolidation and re-organisation of EMA working parties will be required to ensure sufficient expertise and capacity is available for Big Data related applications. More specifically, a consolidation of methods related working parties such as the biostatistics working party, modelling and simulation, extrapolation and pharmacokinetics is envisaged which would be further strengthened by the addition of expertise on RWD and machine learning / AI. The omics technologies remain an area where optimisation is urgently needed. Even in the relatively structured world of genomics, there are multiple regulatory challenges associated with next generation sequencing including lack of standardisation and

Sound implementation of the MDR and IVR and convergence of pharmacovigilance systems is needed

Public Consultation Comments

[16] IPD is defined as data, including imaging data, at an individual patient level which is directly assessable in terms of re-analysis or additional analyses.
[17] https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-qualification-opinion-total-kidney-volume-tkv-prognostic-biomarker-use-clinical-trials_en.pdf

rapidly evolving hardware and software[18], challenges which are amplified for proteomics [6]. To manage the particular challenge of proteomics, formation of an omics working group, building on the foundation of the existing pharmacogenomics group is proposed.

Finally, the last few years have seen an increase in the number of products where integrating gen(omics) biomarkers have become an essential component of medicinal product development and patient selection and thus the performance of the diagnostic test (in vitro companion diagnostic) impacts significantly on the benefit-risk profile of the medicine. While lines between these areas of responsibility are becoming increasingly blurred, currently the assessment of the performance of the diagnostic test is conducted by notified bodies and thus disconnected from the benefit-risk assessment of a corresponding medicinal product. It is also difficult, if not impossible to apply a total life cycle approach to advanced medical devices (especially those relying on bioinformatics algorithms, ML / AI algorithms) when approval and surveillance is carried out by different regulatory bodies, with notified bodies being private companies. The new *in vitro* diagnostic regulation seeks to address this issue but will require the establishment of systematic ties between the medicines regulators and notified bodies to be able to consistently and adequately address such products.

What this means for stakeholders:

By building the capability for regulators to receive, manage and analyse Big Data including PLD from clinical trials, regulators can validate analyses performed by the industry and test assumptions. This will further strengthen the assessment of product dossiers and will underpin benefit-risk decisions. For patients, regulatory validation of study results will provide reassurance that medicines continue to be authorised based on robust evidence.

5.7. Ensure a secure and ethical data sharing culture

Data sharing and access to data fundamentally depends on creating a trusted environment with transparency as regards the intended use of the information and compliance with data protection laws.

> *An ethical framework for Big Data is needed that protects patients over the long-term where they can opt in or opt out as BIG DATA and their own health evolves over time*
>
> *Public Consultation Comments*

In this sense the regulatory context may provide a trusted broker to drive data sharing and it is useful to explore what added value the regulatory involvement may bring.

Big Data brings unique challenges; in a fast-evolving data landscape, where the structure, content and provenance of the data may be unclear the challenge is to enable data sharing while fully respecting data privacy and maintaining the scientific utility of the data. Both data privacy and data sharing can and should co-exist recognising a strong need for transparency, tools and guidance. However, currently there is no concrete data protection code of conduct applicable across Europe which would guide the sharing, linkage and processing of personal information for regulatory use, which are certainly activities performed in the public interest. As highlighted by Vayena and Tasioulas[19] "ensuring that patients fully understand how their data will be used and by whom, and more generally speaking consumer engagement is a key factor for overcoming legal and ethical

[18] https://www.ema.europa.eu/en/documents/report/highlight-report-fourth-industry-stakeholder-platform-research-development-support-23-november-2018_en.pdf
[19] Vayena E, Tasioulas J. 2016 The dynamics of big data and human rights: the case of scientific research. Phil. Trans. R. Soc. A 374: 20160129. http://dx.doi.org/10.1098/rsta.2016.0129

barriers to effective and robust use of Big Data in comparative research, clinical decision support and quality improvement".

In addition, data sharing and secondary use of data for research raises ethical issues which require identification, examination and guidance [7]. In this context "data ethics" is an important consideration. Floridi and Taddeo[20] consider this as a new branch of ethics which "studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing, and use), algorithms (including AI, artificial agents, machine learning, and robots), and corresponding practices (including responsible innovation, programming, hacking, and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)".

Patients as partners in research should always have the right to decide how their data will be used

Public Consultation Comments

The current ethical debate is being dominated by considerations of data privacy but issues exist outside of this, many of which are relevant to the regulatory sector. This may refer to challenges of data "ownership" (especially in a world where data is becoming increasingly commercialised), trust, measures of accountability, bias, group level ethical harms, the distinction between harm to data subjects resulting from respectively academic and commercial uses of Big Data, feedback to participants on unanticipated incidental findings arising for Big Data research and the use of data anonymisation as a mechanism to share datasets which removes the need to request the consent of the data subject. Moreover, the majority of these challenges are not only ethical, but are interconnected with legal and data protection considerations. Ethical principles, conceptual tools or legal requirements for ethics review in clinical research were largely developed in a different era and are now faced with these new challenges.

Lastly, Big Data analyses blur the traditional boundaries between medical specialities, which creates additional challenges for ethics panels. As the European regulatory network begins to consider analyses of Big Data at the patient level and increasingly aims to link these data (especially with sensitive genomic and imaging data), it is important that any potential ethical issues are identified and addressed. Clearly it is critical that patients and healthcare professionals are at the centre of all conversations.

Recommendations

Regulators to engage with data protection initiatives to ensure regulatory use cases for secondary use of healthcare data are understood and to deliver data protection by design

Form an ethical advisory committee to advise on ethical aspects of regulatory use of Big Data. Full patient representation should encompass the spectrum of diseases and age groups

Pilot study to establish patient and health professional views on data sharing including data protection and ethics, among a representative sample of the EU population e.g. across MS, across diseases and across age groups

Promote and support initiatives exploring novel technological solutions to support data protection

Identification and tracking of cases and concerns of the regulatory use of personal data by Member States, EMA Committees and working groups on data protection and data ethics

Fiche #11

The need for guidance on interpretation of European data protection legislation is increasingly recognised and communication between all the relevant actors is required. However multiple initiatives are ongoing (see fiche #9 for details) and hence to avoid

[20] Luciano Floridi and Mariarosaria Taddeo What is Data Ethics? Phil. Trans. R. Soc. A 374:20160112. http://dx.doi.org/10.1098/rsta.2016.0360

duplication of effort, engagement of regulators with existing or planned initiative is proposed in order to ensure regulatory needs are fully considered. This also applies to the development of an ethical framework that can address the ethical challenges associated with Big Data.

An additional route to address personal data protection requirements, and an approved legal basis for data processing, is to anonymise the data whilst maintaining sufficient information to conduct scientific research. As such, data anonymisation can be viewed as an enabler for clinical data sharing while understanding that anonymisation can never be 100% absolute. There is always a residual risk of patient identification which is more likely in a Big Data context as data are triangulated with other datasets of which the data sharer might be unaware, for example those released by other data holders or generated by the individual via social media posts or internet searches. Layered on top of these considerations is the fact that data generation may be single or multiregional and data sharing is likely to be global, and thus these activities must comply with regulations across multiple jurisdictions. As a result, population level uniqueness is increasing as the number of attributes available for an individual increase. Data Protection Authorities have also argued that data anonymisation must be re-evaluated over time as the data environment changes[21]. This challenges the technical and legal adequacy of a release-and-forget anonymisation model, and speaks to a need that re-identification risks should be reassessed regularly. This in itself may limit some of the potential offered by Big Data. Given the second phase of EMA Policy 0070[22] seeks to develop mechanisms to publish PLD while complying with privacy and data protection laws, it is important that the regulatory network continues to explore novel technological solutions to ensure data protection.

Clarity on definitions – anonymised only when all identifiers removed and impossible to re-identify individuals even when triangulated with other data

Public Consultation comment

The potential insights offered by health care data makes the field attractive to commercial companies, increasingly to a small number of large technologies companies, who have the necessary expertise and resources to dominate the market. One approach in this field could be to introduce user accountability into data sharing such that access to data is managed via institutional systems for authentication and authorisation [8]. This may increase patient trust by allowing hosts to enforce proportionate safeguards for datasets that may be sensitive and consented for use only by certain institutes or/and for specific purposes.

Entry of regulators into the big data field is essential

Public Consultation comment

Collaboration and engagement of regulatory authorities with all stakeholders, especially Data Protection Authorities, and where appropriate the European Data Protection Board (EDPB) and the European Commission to facilitate a harmonised approach regarding the use of Big Data is critical to avoid duplication of effort and mixed or inconsistent messages which might hamper Big Data use.

What this means for stakeholders:

By building a governance framework for secure and ethical use of data we ensure that personal data are protected and that ethical challenges are addressed. Ensuring patient and healthcare professional engagement will help overcome legal and ethical barriers to effective and robust use of Big Data in research, clinical decision support and quality improvement.

[21] Opinion 05/2014 on Anonymisation Techniques of the *Article 29 Data Protection Working Party*, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[22] https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf

5.8. Increase network skills via training and strengthening of external collaborations

Our regulatory decisions require a high level of expertise across the scientific spectrum. The current expectation for regulators to keep sufficiently up to date with scientific developments across a wide range of diseases and methodologies in order to deliver critical assessments will become increasingly challenging. New analytical approaches enabled by the availability of Big Data will bring fresh challenges for which the expertise is not currently available in the network. As an example, in 2014, a review by Bauer and Konig [8] highlighted that less than half of the national competent authorities (NCAs) had educated statisticians in their staff. This situation has not improved substantially; in the survey of NCAs conducted during Phase I of the BDTF[23] 8 out of 24 European NCAs who responded to the survey still had no in-house statistical support with the remainder only having between 1 and 5. Unless the regulatory network develops a plan and identifies key initiatives needed to support its work there is a significant risk that sufficient expertise will not be in place to adequately assess such applications and develop the necessary guidelines to advise industry. In addition, support for increasing the expertise of the network is needed to effectively regulate the increasing number of m-health and diagnostic devices and validate the surrogate endpoints these devices may deliver [9].

> *The EU regulatory network needs to develop strong collaborative links with academic institutions*
>
> *Public Consultation comment*

Recommendations

Short-term development of a set of training modules on Big Data topics to meet immediate regulatory needs

Comprehensive skills analysis of the EU regulatory network based on network's strategic needs

Development of high-level training curriculum on Big Data to meet immediate regulatory needs

In the light of a skills analysis EUNTC in collaboration with EMA committees to Develop and implement a long-term big data capacity building training strategy

Establish framework for external knowledge exchange to support training needs

Supplement training by PhD/MSc projects, short-term external placements and targeted recruitment

Promote regulatory science centres as a vehicle to drive collaborative regulatory science projects

Fiche #12 & 13

Ultimately, skill requirements and training needs will be determined by which recommendations of the BDTF are prioritised by HMA and EMA's Management Board. At this point further identification and detailing of the required skills and analysis of the skills gaps is required. Based on the outcome, a full capacity building strategy can be developed detailing a multifaceted approach to closing the detected skills gaps. This includes internal training development coordinated by the EU NTC, but also other capacity building approaches such as knowledge exchange with academic institutions, PhD and MSc projects both internal and external to regulatory agencies, short-term placements or sabbaticals and collaborations with regulatory science centres as a vehicle to drive collaborative science projects.

> *Strengthen training on 'omics – there is a lack of expertise on how to study/read the large amount of data produced by 'omics studies*
>
> *Public Consultation comment*

[23] https://www.ema.europa.eu/en/documents/other/hma/ema-joint-task-force-big-data-surveys-results_en.pdf

As an interim measure to meet immediate regulatory needs, a pragmatic set of training modules on topics related to Big Data in regulatory science has been proposed for short-term development (coordinated by the EU NTC). This is detailed in Fiche #10.

A long-term strategy to deliver capability and capacity sustainably will require a multifaceted approach involving knowledge exchange with academic institutions.

What this means for stakeholders:

By building the expertise available to the EU regulatory network industry can get comprehensive scientific advice to guide their selection of data source and analytical approach. Regulators will be equipped to assess authorisation submissions based on novel datasets and be able to seek confirmatory evidence based on a detailed knowledge of the data and methods.

5.9. Drive continual optimisation of the regulatory assessment of Big Data approaches

Regulators need to use data to make reliable, reproducible, scientific and evidence-based decisions. This need may not always sit well with the dynamic accumulation of data and refinement of approaches in the Big Data field. In addition, more and more post-authorisation, ad hoc assessments of the benefit-risk of products by academia, civil society and activists will challenge both companies and regulators alike.

The process of optimisation can be arbitrarily divided into optimisation of data and processes and of the analytical approach. In the Big Data field there is often uncertainty around the quality of the data and while characterisation of the data via a data quality framework will inform on the current status it will not necessarily drive improvement. Current standards and specifications need to be continually enhanced and new ones developed as innovation and technology advances to provide clear, simple rules and guidance for all stakeholders. Moreover, with the increasing digitisation of data capture, healthcare records and even healthcare delivery, comes an increasing need to align international data standards to improve data interoperability and methodology to support evidence generation. Development of a clear data standards strategy combined with a development framework and action plan would do much to enhance interoperability of European healthcare and health related data sources and support more efficient pre-and post-authorisation review at the PLD level.

> *The regulatory system should be prepared for and understand the change in data generation and knowledge management*
>
> *Public Consultation comment*

Use of a federated approach to analysis across multiple datasets holds promise. Data remains where it is and the probe arrives with the question, runs the algorithm and returns an (anonymised) answer at an aggregate level. This would in principle overcome many of the data privacy concerns that Big Data leads to as the technology doesn't require data to be moved. Distributed/federated data models do have the drawback of losing some statistical power compared to a model with one common data repository, but even small degrees of data optimisation (standardisation of the distributed data towards a common set of accepted standards) might ameliorate this. Deep Learning (see section 5.10) models could interrogate distributed data sets given appropriate data access is in place. Alternatively, a common data model (CDM) can be used to push out analytic code, have the branches of the federated structure run the analysis which can then be compiled with meta-analytic approaches for a summary effect measure.

With regard to companion diagnostics for medicinal products and devices that measure biomarkers in clinical trials, we are dealing with two different legal frameworks. This may not apply for every source of Big Data but for those areas where a 'device' is used e.g. in the capture of clinical-trial data or in the analysis of clinical data, or the detection of biomarkers, it is important that the two parallel worlds work together. For example, the regulation of next generation sequencing as a diagnostic tool will fall under the in vitro diagnostic regulation and thus standards will be set under this framework to allow them on the market. There is clearly a need for cross-collaboration and input to set one standard (if applicable).

Key to continual optimisation is effective engagement with stakeholders (industry, government, academia, healthcare professionals and patients) and a convening platform to enable consultation and the establishment of mutually beneficial pilots would accelerate progress. Recent examples of consortia/focus groups include post licensing evidence generation[24] and the digital therapeutics alliance[25] but a broader platform addressing a range of issues is envisaged here which would include patients, healthcare professionals and academia. In the context of a fast-moving scientific landscape there is a need to improve the agility of guidance development. Mechanisms need to be found to provide more agile advice possibly via Question & Answer documents, discussion documents to reflect interim thinking and stakeholder engagement through conferences.

> *'Support a data analytics group to explore the application of technologies and encourage continued stakeholder engagement to inform subsequent regulatory guidance'*
>
> *Public consultation comment*

Recommendations

Establish a stakeholder consultation platform (industry, government, academic and patients) to address Big Data related questions and processes

Seek international regulatory alignment on data standards, data interoperability and methodologies for Big Data studies

Formation of a big data steering group for oversight, horizon scanning and to agree data science research priorities

Seek alignment and harmonisation on regulatory requirements with all actors in the European health sector

Develop new models to increase the agility of guidance development

Establish a federated network of advanced analytical centres

Fiche #14, 15, 16, 17

Consistent horizon scanning should be implemented to ensure that regulators keep track of externally funded research initiatives (both at a national, European and international level) which may potentially impact on or support regulatory decision-making. While the EMA maintains close contact with, in particular, IMI funded initiatives, a centralised knowledge of nationally funded initiatives is more patchy. This is key as an assessment in funding for RWD initiatives over the last decade revealed the immediate utilisation of the outputs of European funded initiatives to support regulatory decision-making is limited, often due to insufficient availability of information, and to discrepancies between outputs and objectives [10]. Multiple projects focussing on the same therapeutic areas increase the likelihood of duplication of both efforts and resources. Furthermore and importantly, the restricted sustainability of the majority of these initiatives significantly impacts on their downstream utility. These issues contribute to gaps in generating RWE for medicines and diminish returns on the public funds invested.

[24] https://www.ema.europa.eu/en/documents/report/highlights-report-second-industry-stakeholder-platform-research-development-support_en.pdf
[25] https://www.ema.europa.eu/en/documents/report/highlight-report-fourth-industry-stakeholder-platform-research-development-support-23-november-2018_en.pdf

The rapid expansion of analytical approaches, principally AI, has introduced the possibility of using Big Data sources to develop new models to deliver healthcare e.g. automated diagnosis, disease stratification and in the regulatory world provide a more personalised approach to medicines, which may ultimately personalise not only dose and benefit, but also risk. However, there are many challenges. Firstly, we need to understand how to assess the evidence generated by our stakeholders using these new methods; currently there are no definitions of acceptable performance standards, accuracy, disease areas and patient health outcomes against which to measure AI. Secondly, to determine whether AI can work with the data provided or whether it will always require additional input or actions to improve it. Thirdly, to create guidelines on the use and validation of new analytical techniques, especially on the generalisability of Big Data insights derived from AI based predictions on country specific data to other countries and healthcare systems. Lastly, to generate the expertise within the regulatory network in this growing area where currently there is limited capacity.

Better horizon scanning to identify where existing data can answer new questions

Public Consultation comment

The BDTF recommendations will likely influence the Network Strategy to 2025, the EMA Regulatory Science Strategy to 2025 and the individual work plans of EMA committees and working parties. However as no further phases of the BDTF in its current form is foreseen, a Big Data/analytics steering group is proposed to ensure success criteria are agreed and monitored, to maintain oversight of the implementation of ongoing and proposed work, to ensure continued horizon scanning of the scientific landscape and to agree on data science research priorities for the regulatory network.

What this means for stakeholders:

By engaging with stakeholders in the EU and beyond we can share best practice avoid duplication of effort, effectively horizon scan and ensure the voice of our stakeholders is heard. This will ensure medicines regulation remains relevant and fit for the future. A Big Data Steering Group reporting to HMA and EMA Management Board will ensure accountability for the implementation of key recommendations of this report.

5.10. Regulatory considerations on Bioinformatics, Algorithms, Machine Learning and Artificial intelligence (AI)

With the development of precision medicine the use of technologies like genomic sequencing and the use of the required bioinformatics tools and algorithms is well on the way, or in some cases already established. Numerous genomic markers have been tested and used for patient stratification in clinical trials. In case bioinformatics tools and algorithms are used in the context of companion diagnostics, regulatory agencies are directly concerned. It has to be ensured that the European regulatory network has the expertise and capacities in order to asses these applications appropriately.

Beside the use for diagnostic purposes, these innovative technologies can foster the development of innovative medicines in a broad variety of applications. For instance, in the growing field of therapeutic vaccines there are different approaches and strategies under development. In some cases, these development candidates are already in late clinical phases. Bioinformatics tools and algorithms are i.e. used to identify mutation-containing epitopes that are predicted to bind to the MHC class I molecules of individual patients. These bioinformatics tools determine the individual medicinal products, like the sequence of distinct peptides, individually synthesised for each patient. Thus, the manufacturing and subsequently, the safety and efficacy of these medicinal products is entirely dependent on the correct and robust functionalities of the bioinformatics algorithms used.

AI, defined as a self-learning evolution of well-known adaptive statistics, is already part of the regulatory landscape. ML algorithms where the algorithm incorporates feedback to continuously optimise output has been incorporated in randomisation algorithms and many devices use these techniques. Further development of AI into Natural Language Processing (NLP) recognizing and processing free text recognising images, robotics guiding surgical instruments, and even Deep Learning Algorithms (DL) are now part of the data handling landscape regulators have to assess.

These techniques, where implemented into clinical trials have led to the understanding that efficacy of such technologies strongly depends on the understanding of the practical situation under investigation and will not universally lead to an improvement over more standard techniques. Such experiences underline the need to fully understand and carefully validate new strategies for decision-making when based on complex algorithms.

> *'Cognitive computing offers new capabilities to add speed, scale and consistency to the entire pharmacovigilance process from adverse event intake, triage, evaluation and reporting, to signal detection and assessment'*
>
> *Public consultation comment*

The European regulatory system must continue to ensure safe and effective medicines for the European Public, therefore knowledge of these new techniques must be incorporated into the regulatory network as they will directly influence clinical decision-making and thus public health.

Four immediate areas of AI use are important to address in a regulatory context:

1. Ensure sufficient expertise and capacities are available within the European network, in order to assure that bioinformatics tools and (ML- or AI- driven) algorithms can be assessed appropriately if these technologies are used in the context of regulatory submissions.
2. Enable regulatory evaluation of clinical data submitted by drug manufactures for approval where the data has been processed by AI algorithms or part of the analysis, such as patient selection, involved AI methods.
3. Explore regulatory use of AI in internal processes – e.g. NLP processing of text – categorizing eCTD submissions into review templates for assessors or quantitative review of image data submitted to support a clinical claim from a drug manufacturer.
4. Approval of AI-based Health Apps in devices intended for clinical decision-making.

Several considerations become apparent when one considers the constantly evolving nature of AI Firstly, no algorithm will ever perform any better than the data sets it is trained on. Thus, the output will reflect the distribution and variability of data in the training data set. For example: if an algorithm is only trained on Caucasian western European data it may not be very predictive on outcomes for southern-/eastern European populations or immigrants to Europe of African descent. It is therefore imperative that regulators require algorithms to 'explain themselves' i.e. to be programmed in advance with a view towards being interrogated. Also, AI algorithms need to flag data outside the distribution of the training data set as these may not be accurate predictions.

It is clear that regulators cannot and should not accept the so-called 'black box' concept where algorithms simply perform in a vacuum without any checks and balances. Algorithm code should be more transparent (feature selection, code, original data set) and available for targeted review by regulators. Outcomes of and changes to algorithm use (safety and efficacy) needs to be subject to post-marketing surveillance mechanisms, just like it is done today to monitor drug safety after marketing authorisation.

Employing bioinformatics tools and algorithms, ML and AI must still be on secure, transparent, reliable and reproducible data. For instance, it may be difficult to use established validation approaches to AI technology but in several areas it is possible to validate against gold standards e.g. in quantitative imaging analysis where measurements done by trained radiologists on DICOM-standard images can serve as the standard. EMA Qualification Advice is considered as a suitable tool to address and evaluate the use these innovative data analysis tools for regulatory purposes.

> *'Does the HMA/EMA plan on attending/presenting at industry/academia conferences and/or providing webinars and updates on their position on Big Data as it develops and is implemented?'*
>
> *Public consultation comment*

There is no doubt that the use of bioinformatics tools and algorithms, ML and AI- has the potential to greatly improve the development, manufacturing and application of medicines, as well as the data handling and even accuracy and predictability in health care. But it must be held accountable to regulatory standards - it is imperative that regulators and decision makers now invest in upskilling their staff and the regulatory infrastructure to meet these new challenges. Only then can the true potential of these technologies be safely deployed.

Last but not least, a regulatory network, holding relevant expertise and with access to required capacities and resources, is essential to ensure an innovation friendly environment, enabling European researchers and developers to play a leading role while establishing the use of these innovative technologies in health care – for the benefit of citizens and patients.

One way to handle this apparent lack of analytical resources would be to create clusters of expertise, i.e. NCAs with sufficient computational infrastructure and analytical expertise to perform the necessary analyses on behalf of the European regulatory network.

What this means for stakeholders:

Establishing centres for analytics, the ability to validate algorithms and the processes to enable advanced analytics of healthcare data holds the promise to facilitate the development of innovative, often targeted medicines that can fulfil the unmet medical needs of patients.

5.11. Develop an overarching strategy to communicate regulatory approaches in the Big Data field

Many diverse recommendations are proposed by the BDTF, which will impact on multiple stakeholders in an area which is not only complex with many complicated and ill defined concepts but which incorporates discussion of ethically sensitive issues. Thus, an overarching communication strategy is needed to ensure coordinated external communication with consistent messaging; the strategy should deliver a balanced description of Big Data activities which highlights not only the benefits at a population level but also discusses the potential risks to evidentiary standards. Clear and transparent communication will aim to build trust in data sharing exercises and external support for specific BDTF and regulatory initiatives by communicating the public health benefits as well as describing the robust governance. Moreover, it is important to optimise communication to support the other activities of the BDTF. The communication strategy should also proactively identify and mitigate negative or erroneous perceptions that risk the achievement of the vision. During the

> *'Patients should not be viewed as data generations but as decision makers empowered to drive big data insights in their own right'*
>
> *Public consultation comment*

implementation of any recommendations, it will be important to develop metrics for reach and impact, including quantitative (e.g. web statistics, shares, media mentions) and qualitative (e.g. positive impact on an influencing activity, stakeholder survey, citations, uptake of BDTF recommendations) metrics and the assessment of activities such as membership of standards organisations and presentation at conferences.

Recommendations

Proactive external communication to raise awareness of regulatory needs as well as the outputs of the BDTF

Support patient awareness on the need of systematic collection of information on disease, treatments and outcomes

Define metrics for reach and impact

Fiche #18

What this means for stakeholders:

Clear articulation of the regulatory needs and use cases for Big Data will enable stakeholders to optimise their role in realising the potential of these data for public health.

5.12. Big Data Initiative: Data Analysis and Real World Interrogation Network (DARWIN)

Regulatory gap

To adequately and proactively monitor the benefit-risk of medicines across Europe, potentially for decades after the initial treatment point, the European regulatory system requires timely access to data representative of the whole of Europe and of sufficient quality and scope to support its decision-making. The ability to reproduce RWD analysis across multiple data sources will reassure decision makers of the robustness of the evidence

While not the only relevant data source, it could be argued that RWD has the most immediate potential to address additional evidence needs across the product lifecycle [11]. Until recently accessing such data at scale was not technically possible. However, over the last decade, technological advances have opened up new possibilities to access and analyse multiple complementary data sets and increasingly these can bring real value to regulatory decision-making.

Currently several NCAs access, either directly or indirectly, their national healthcare databases to inform decision-making. Similarly, the EMA is routinely using three RWD sets and over the last 5 years has performed 72 in-house studies to directly support the evidence needs of EMA Committees, mainly the PRAC. In addition to these in-house studies, EMA has directly commissioned, via a network of academic centres across Europe, 15 external studies, most of them multi-database and multinational. While this approach delivers reassurance by enabling the replicability of study results over multiple databases, it takes time for procurement exercises, in securing academic time, agreeing protocols and securing data access across multiple sources and hence delays decision-making. It is also largely limited to general practice data and with limited geographical spread across the Member States.

While it is clear that analysis of RWD databases can provide valuable information to support regulatory decisions [11], [12] , there are major challenges associated with their use which many of the BDTF recommendations seek to address. It is important to recognise that the challenges require concerted

action, and that they are not challenges that the regulatory network can solve alone. While the need is well recognised, and efforts are underway, they are slow and inconsistent across Europe and moreover the regulatory system is not well engaged with these efforts in order to articulate regulatory needs.

Key among all these challenges is access to data to enable its efficient and timely use throughout a regulatory assessment process. While Europe is fortunate in the richness of its healthcare data, and in particular its longitudinal nature, which stems from the principle of universal healthcare coverage, it lacks the means to fully exploit it. This is partly because European healthcare data are heterogeneous as differences in healthcare systems, national guidelines, and clinical practice have driven different content, and historically different healthcare systems have used different terminologies and structures. However, it is also partly a result of a lack of focussed and sustained funding which means that, despite the investment of more than 734 million Euros into 65 healthcare related initiatives by EU centralised funding vehicles [10], as a regulatory system we do not have systems to match those at the disposal of other regulators. In addition, other international regulators have committed significant resources to establish networks of distributed real-world databases alongside strategies to enable routine access for regulatory purposes e.g. Sentinel (FDA), CNODES (Health Canada) and MIDNET (PDMA). For example, Sentinel now provides the FDA with access to 100 million unique patient identifiers[26], and is constantly growing and evolving, incorporating new data partners to address some of the limitations of US healthcare data. To manage and drive its development FDA devotes significant internal resource, both financial and human, and early on appointed a co-ordinating/operational centre to manage and maintain the data sources and run studies. Furthermore, to facilitate interrogation of the databases in a timely manner, Sentinel developed a toolkit to enable standardised analyses of increasing complexities from simple descriptive analyses (level 1) to adjusted analyses with sophisticated confounding control (level 2) and finally the most sequential adjusted analyses with sophisticated confounding control (level 3). Yet further development is envisaged in the Sentinel System five-year strategy including increased data granularity, the broader use to evaluate effectiveness and the use of more sophisticated analytical techniques including data mining capabilities[27]. However, such sustainable solutions require the reassurance of funding to secure the basic functioning of any data network. This has been demonstrated by the FDA Sentinel network which requires baseline funding of $10,000,000-$15,000,000 per annum and other European initiatives such as UK Biobank which receives charity and governmental funding to secure its activities.

Recommendations

Proactive external communication to gain stakeholder support for DARWIN

Develop a clear business case which defines the delivery and sustainability model for DARWIN, including financial needs

Establishment of an analytical system to support real time analysis of DARWIN

Development of regulatory use cases to inform thinking and ensure DARWIN is fit for purpose

Fiche #18

As a regulatory network and in the Big Data era we must address this and hence a key recommendation is the establishment of a European network of databases of known quality and content who agree to the highest levels of data security, which the BDTF calls DARWIN: the **D**ata **A**nalysis and **R**eal **W**orld **I**nterrogation **N**etwork. The network should be scalable to allow for different speeds of adoption and implementation.

[26] If a patient moves health plans they may have more than one patient identifier. Currently it is not possible to link patients between different health plans.
[27] https://www.sentinelinitiative.org/communications/publications/sentinel-system-five-year-strategy-2019-2023

This recommendation has many components and poses many operational and technological challenges. Broadly, these can be divided into operational, technical and methodological challenges; as described in a recent review (OPTIMAL) [13], addressing these as a whole requires the development of a framework for regulatory use of real-world evidence. Some of these challenges will be addressed with specific recommendations elsewhere in this report but it is clear that in order to move the dial and bring all stakeholders together, including patients and healthcare professionals, a clear business case for DARWIN needs to be developed. This should set out among other aspects the strategic case, background and current landscape including ongoing relevant activities, key objectives, relevant stakeholders anticipated benefits and success criteria, key risks and proposed mitigations as well as interdependencies, assumptions and constraints. It should define potential funding sources; securing significant EU set up funding (likely of the order of tens of millions) and long-term sustainable funding of approximately EUR 10-20,000,000 per year would could potentially be delivered through a new dedicated fee implemented through the proposed revision of the EMA fees regulation. The business case should include a delivery plan along with clear timescales and sustainability model for DARWIN. Annex 0presents a draft business case for DARWIN.

6. Resources

The mandate for Phase II of the Big Data Task Force included both prioritisation of recommendations from Phase I and high-level planning including resources needed from the EU regulatory network, to guide decisions on implementation.

The recommendations of the BDTF can be summarised in terms of types of output, as follows:

- 1 big collaborative initiative to access and analyse health data: DARWIN.
- 1 Steering Group to guide implementation of the BDTF recommendations.
- 1 cross-committee task force on practicalities of receiving and analysing PLD.
- 1 stakeholder implementation forum, together with workshops on specific topics within Big Data.
- 2 Networks of centres of excellence (analytics and regulatory science).
- 2 Processes (qualification advice and agile guidance development).
- 3 reformed or new working groups (methods, omics, ethics).
- 4 IT domains (data management capability, analytics capability, ENCePP resources database, EU PAS database).
- 6 Guidelines.
- 6 Pilots (data quality framework, hackathon on AI use in EudraVigilance and ADRs linked to genomics, RWE in drug development, PLD from clinical trials, patient views on data sharing).

It is suggested that certain principles should guide the implementation of the BDTF recommendations:

- Collaborate: external stakeholders.
- Require: good practice from the industry we regulate.
- Network: with our NCA partners to deliver Medicines Network solutions.
- Protect: the best parts of our current system e.g. clinical trials for efficacy.

- Build: on what we are doing e.g. bring together working parties on methods.
- Recognise: the excellence we have and train our staff.
- Target: recruitment for specialist skills.
- Leverage: planned stakeholder initiatives e.g. EC funding for digitalisation.
- Seize: opportunities as they arise e.g. revision of the EMA fee regulation, EMA Regulatory Science Strategy.
- Collaborate: to deliver more e.g. ICH, ICMRA and bilaterally with FDA, HC, PMDA.

For the detailed assessment of resource requirements, a three-step approach has been employed, as follows:

- Step One: BDTF sub-groups have estimated resources when developing fiches for specific deliverables (bottom up).
- Step Two: the BDTF co-chairs have estimated resources for all the recommendations included in the final report (top down).
- Step Three: at the plenary BDTF meeting on 9 October 2019 the estimates from Step One and Step Two have been compared and clarified.

Given that detailed project plans have not been developed for each individual recommendation, resource estimates should be considered as being 'rough order of magnitude' and are included to support future planning. Annex 9.7. presents a breakdown of estimated resource implications for the EU regulatory network, broken down by individual recommendations. Estimates are further organised as:

- Staff (Full Time Equivalents - FTEs)
- IT cost (per year ROM – T-shirt approach)
- Meetings/delegates costs (number of meetings / workshops per year)
- Missions (either EU or non-EU per year)
- Consultancy (costs per year)
- Other costs.

It should be noted that to respect the principles that we should: build on the existing EU network organisations and structures; build on what is already working well; and, to avoid duplication of counting, the BDTF resource estimates for the network have not included better use of existing resources where no significant new cost is incurred. For example, network staff who currently work on the assessment of marketing authorisation applications will continue to do so in the future even if many of them will have undergone further training in statistics and epidemiology and on the spectrum of different EU data sources.

Acknowledging the principles and caveats described above and particularly that the opportunities of Big Data will be fully realised through collaboration with stakeholders, minimum direct costs to the EU regulatory network (EMA and individual National Competent Authorities) if all the BDTF recommendations are implemented are summarised here.

New staff costs (FTE) not cumulative:	Working groups and workshops (number of meetings):
• Year 1: 24	• Year 1: 16
• Year 2: 24	• Year 2: 28
• Year 3: 19	• Year 3: 23

Missions	
Within the EU	**Outside the EU**
• Year 1: 40	• Year 1: 24
• Year 2: 43	• Year 2: 25
• Year 3: 50	• Year 3: 20

Consultancy (Euros)
• Year 1: 500,000
• Year 2: 500,000
• Year 3: 500,000

Information technology costs (Euros)
IT costs Year 1: 1,320,000 – includes:
• set up of secure cloud architecture for data integration
• software for rapid analysis of EHRs
• start upgrade of the EU PAS Register
• start upgrade of the ENCePP Resources Database
• set up hackathon infrastructure and run proof of concepts
IT costs Year 2: 1,560,00 - includes:
• maintenance of cloud architecture for data integration (supports pilot of network analysis)
• maintenance of rapid analysis of EHRs
• infrastructure and software for the pilot to analyse and visualise patient level data (PLD)
• software for AI analysis
• complete upgrade of the EU PAS Register
• complete upgrade of the ENCePP Resources Database
• maintenance hackathon infrastructure and run hackathons

Information technology costs (Euros)
IT costs Year 3: 1,880,000.00- includes:
• maintenance and scale up of cloud architecture for data integration (supports pilot of network analysis)
• maintenance of rapid analysis of EHRs
• maintain software for the pilot to analyse and visualise some patient level data (PLD)
• maintain software for AI analysis
• maintain the EU PAS Register
• maintain the ENCePP Resources Database
• maintain hackathon infrastructure and run hackathons

Other annual costs (Euros):
Year 1: 1,050,000.00
• Access to EHR data
• Data Quality Framework
• Genomics proof-of-concept
• Training
• Translations
Year 2: 1,750,000.00
• Access to EHR data
• Data Quality Framework
• Genomics proof-of-concept
• Training
Year 3: 1,750,000.00
• Access to EHR data
• Data Quality Framework
• Genomics proof-of-concept
• Training

7. Conclusions

The previous report of the BDTF set out a number of recommendations to address what are well-recognised challenges if Big Data is to deliver evidence of suitable strength to support decision-making across multiple stakeholders. Based on these a prioritisation was required to focus in on those changes which will best prepare our regulatory model for future challenges. Therefore, this report "Evolving Data-Driven Regulation" represents the final deliverable of the HMA-EMA Joint Big Data BDTF.

It must be emphasised that Big Data in itself is not necessarily the solution to all the challenges faced by regulators in reaching robust decisions, but the complementary evidence it generates will facilitate, inform and improve our decisions. However, despite all the hope, the tangible outputs of Big Data in the context of regulatory decision-making still sit closer to aspiration than reality. Nevertheless, it is clear that the data landscape is evolving and that the regulatory system needs to evolve also and to prepare for and understand the diversification in data generation and knowledge management that will be required.

In this report, in order to make concrete progress and move closer to reality, in its second phase of work the BDTF has focused on our regulatory processes identifying initiatives which would strengthen our regulatory paradigm and which therefore have the potential to impact more immediately on public health. Such initiatives, focussed on how we can understand data better and the evidence generated from it, are urgently needed to accelerate the translation of innovation and research findings through to new safe and effective medicines for patients as early as possible. However, such evolution does not come for free; the regulatory network, society and stakeholders will need to devote resources in a sustainable way to move to new ways of working which incorporates new technologies and datasets. Our knowledge management processes will also need to evolve and investment is needed to build the necessary expertise to enable us to appropriately guide drug development. Resources have been suggested for the European network for the immediate implementation phase but in addition it will be important to engage proactively with external funding bodies to capitalise on the current interest in digitisation and ensure that healthcare and regulatory needs figure predominantly in new European initiatives.

Clearly, we must not desert well proven, robust regulatory models designed to eliminate bias in decision-making, but equally we need to strengthen and adapt our currently regulatory model so we are able to confidently extract value out of the data to address the assessment challenges ahead. It is clear that the data landscape is evolving and that the regulatory system needs to evolve also. In this way we can realise opportunities for public health and innovation through better evidence for decisions on the development, authorisation and on-market safety and effectiveness monitoring of medicines. As healthcare data and technology evolve then so must medicines regulation.

HMA-EMA joint Big Data Task Force: What this means for stakeholders

If the recommendations of the Big Data Task Force are prioritised by the EU regulatory network and stakeholders then, already in 2020, the capacity of the network to advise on, assess and analyse Big Data will start to increase. Training will be rolled out across the network, expert working parties will be rationalised and guidance will start to be developed and consulted upon.

By 2023:

A collaboration across stakeholders starts delivering access to and analysis of healthcare data from across the EU (DARWIN);

Data will be discoverable (through the ENCePP resources database) and of known quality and representativeness allowing to choose the optimal data source to generate evidence and enabling regulators to expertly assess study results for robust benefit-risk assessment;

EU Network staff will be trained and have knowledge and experience in data science, 'omics, methods and analytics to advise companies developing products and to expertly assess application dossiers. Committee decision-making will be enriched with expert advice across the spectra of Big Data types and on analytic approaches;

Built on observation and learnings from submissions of Big Data to the regulator and supported by enhanced study transparency (EU PAS Register), a suite of EU and international guidelines and standards will be available to help industry and regulators develop and supervise medicines;

The EU network will be scaling up its computing capacity to analyse Big Data including targeted patient level data from clinical trials, 'omics, analysis of real world data both pre-and post-authorisation and regulatory validation of artificial intelligence algorithms use in products, product development and regulatory processing (including ADR reporting);

Data submitted and analysed in the EU for regulatory purposes will be managed in full compliance with data protection legislation and respectful of patient and healthcare professional concerns on the ethics of data sharing;

Collaboration from bench to bedside will be established built on open two-way dialogue including a stakeholder implementation forum.

By delivering the vision of a regulatory system able to integrate Big Data into its assessment and decision-making, we can support the development of innovated medicines, deliver life-saving treatments to patients more quickly and optimise the safe and effective use of medicines through measurement of a products performance on the market.

8. References

[1] [Online]. Available: NEJMra1510069.pdf.

[2] "Another major drug candidate targeting the brain plaques of Alzheimer's disease has failed. What's left?," Science | AAAS, [Online]. Available: https://www.sciencemag.org/news/2019/03/another-major-drug-candidate-targeting-brain-plaques-alzheimer-s-disease-h.

[3] A. Pacurariu and et al, "Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation," *BMJ Open 8, e023090,* 2018.

[4] M. Ienca and et al, *Considerations for ethics review of Big Data health research: A scoping review.,* PLOS ONE 13, e0204937, 2018.

[5] M. D. Wilkinson and et al, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data,* no. doi:10.1038/sdata.2016.18, 2016.

[6] V. Lampasona and et al, "Islet Autoantibody Standardization Program 2018 Workshop: Interlaboratory Comparison of Glutamic Acid Decarboxylase Autoantibody Assay Performance," *Clin. Chem. 65, 1141–1152,* 2019.

[7] M. Fiume and et al, "Federated discovery and sharing of genomic data using Beacons," *Nat. Biotechnol,* no. 37, p. 220–224, 2019.

[8] P. Bauer and F. König, "The risks of methodology aversion in drug regulation," *Nat. Rev. Drug Discov,* no. 13, p. 317–318, 2014.

[9] C. Schuster Bruce, P. Brhlikova, J. Heath and P. McGettigan, "The use of validated and nonvalidated surrogate endpoints in two European Medicines Agency expedited approval pathways: A cross-sectional study of products authorised 2011-2018," *PLoS Med. 16, e1002873,* 2019.

[10] K. Plueschke, P. McGettigan, A. Pacurariu, X. Kurz and A. Cave, "EU-funded initiatives for real world evidence: descriptive analysis of their characteristics and relevance for regulatory decision-making," *BMJ Open 8, e021864,* 2018.

[11] H. Eichler and et al, "Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?," *Clin. Pharmacol. Ther,* no. 105, p. 912–922, 2019.

[12] L. Li and et al , "Metformin use and risk of lactic acidosis in people with diabetes with and without renal impairment: a cohort study in Denmark and the UK," *Diabet. Med. J. Br. Diabet. Assoc,* no. 34, p. 485–489, 2017.

[13] A. Cave, X. Kurz and P. Arlett, "Real-World Data for Regulatory Decision Making: Challenges and Possible Solutions for Europe," *Clin. Pharmacol. Ther,* no. 106, p. 36–39, 2019.

9. Annexes

9.1. Annex I: Stakeholder responses

- Data Standards

There was significant support for the need for setting data standards across all datasets and an acknowledgement of their key importance in enabling interoperability of data. The fundamental principle of minimisation was supported, that is, the development of a limited number of global, clear, transparent but also flexible standards. Nevertheless, there is a huge implementation challenge especially in Europe where numerous standards and terminologies are utilised. Even where the same standard is utilised there can be variability of interpretation and local modifications and additions to that standard hindering interoperability. These challenges are widely recognised. In terms of clinical data, support for CDISC standards was voiced and in particular the potential role of CDASH[28] for collecting clinical data. It was also noted that format of processed data following data sharing exercises also needs to be standardised.

As a first step an inventory of all data standards currently in use or being developed was proposed to track developments in a very complex landscape of multiple standard organisations and various consortia on a background of a fast-moving scientific landscape. Interaction and representation of regulatory agencies on key standards organisations was encouraged.

- Data sharing/accessibility

Comments principally focused on the mechanisms needed to support and incentivise data sharing. Suggestions included: appropriate acknowledgements for data generators; addressing incentives for academics who otherwise delay sharing in order to exhaust analytic possibilities; creation of quality metrics linked to organisational culture to reward data sharing; robust governance of not only the data but also the results generated through data sharing exercises; managing fears around publication of conflicting results, based on a secondary analysis, compared with the original; controlled sharing mechanisms; and data controllers provided with the opportunity to review protocols for secondary analysis and the subsequent reports. However, realism is required and caution was expressed over the ambition of full data sharing and whether this was realistic. Lastly, the need to ensure sustainability of data sharing efforts was emphasised as was the need to incorporate the patient perspective. Fundamentally, in order to garner support the benefits of data sharing need to be consistently publicised: show casing real case studies to illustrate the benefits of controlled anonymised data sharing.

- Data linkage

The importance of linkage of genomic and 'omic data in general to phenotypic data was strongly supported. A plea was made for a stronger recommendation in this space and in particular for linkage across different sectors and countries.

- Data quality

The key need for an explicit definition of data quality was completely supported but the challenge of defining this in the context of its use across multiple stakeholders was recognised. A certification model is required and it is likely that the current EMA qualification advice process will require tailoring and likely significant expansion, in order to ensure that provided advice will meet the demands. This

[28] CDASH establishes a standard way to collect data consistently across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into the Study Data Tabulation Model (SDTM)

applies for both, novel Big Data methods and strategies, as well as for data sources, such as registries. In relevant cases, repeated qualification processes may be needed as quality may not be a static and re-qualification would be required after major updates. The concept of an independent certification process for data quality and data curation processes should be explored and established.

- Data analytics

Recommendations in the BDTF reported around data analytics were at a preliminary phase as the subgroup was yet to report. Comments highlighted the potential of artificial intelligence analytical technologies to improve the speed, scale and consistency of signal detection and assessment in pharmacovigilance, in particular with regard to unstructured case narratives. However, comments struck a cautious note by also emphasising the risk of being misled by spurious, non-causal associations if epidemiological concepts of confounding and bias were not sufficiently considered; more emphasis on mechanisms to manage these issues is required. In this context, data analysis plans are currently not sufficiently clear; the ability to access at least a sample of the data (or a synthetic dataset) as well as the code used for the studies would reduce duplication of effort and promote efforts on consistency of reporting. A specific challenge will be delivering transparency around the analysis in model free machine learning algorithms prior to the collection of data. A further specific recommendation was that the BDTF should consider federated machine learning solutions which can exploit decentralised data and meet concerns over privacy. A lack of expertise meant recommendations around imaging data could not be proposed by the BDTF; however stakeholder comments urged that this field should be prioritised going forward given the significant advances in computer aided evaluation of images.

- Regulatory acceptability

The general area of understanding when data would be considered acceptable for regulatory decision-making received significant focus in stakeholder comments. Many comments centred around the need for clear guidance on regulatory expectations and acceptability on utility of data across the product life cycle. A framework is needed with guidance on which factors should be considered and addressed within a regulatory submission and comments highlighted the lack of guidance compared with that provided by FDA. However, it is acknowledged that acceptability will always require contextualisation and tailoring for the specific product life style. A number of specific suggestions were made which may support decisions on acceptability: registration of protocols in a publicly accessible and centralised European registry, consideration as to how randomisation through linked RWD could enhance regulatory acceptability of RWE, ensuring incorporation of study design considerations into any framework on RWE acceptability, and regulatory policing of compliance with any framework. Finally, specific recommendations of the acceptability of pragmatic randomised controlled trials would be welcomed.

- Data protection and ethics

The summary report acknowledged that in the first phase of the BDTF data protection and ethics were not considered explicitly but that these fundamental issues would be incorporated into the next phase of its work. Nevertheless, the importance of data protection and ethical, responsible use of patient data to enable data sharing, an obvious requirement for the opportunities of Big Data to be realised was highlighted multiple times. Many of the comments centred around and emphasised the need to liaise with and consult patients and their representative bodies to understand views. This is even more critical in the era of wearables and social media, constantly collecting and processing data in a context where consent forms prepared by commercial companies are far from transparent. In addition, the need for ethical frameworks in addition to addressing the legal basis of data processing was mentioned. Ethical frameworks should develop processes to protect patients over the long-term allow

opt in and opt out of data sharing as not only the Big Data field evolves but also their own health status changes. Examination of consent processes should be a key priority of such a group in particular a consideration of opt in/opt out policies for secondary research, how compliance with policies would be ensured and what policies and processes should be followed when data sets are shared on an anonymisation basis when it is recognised that anonymisation can never be absolute.

- Training and skills

In line with the summary report, it was recommended that the EU regulatory network needs to develop strong collaborative links with academic institutes especially in areas where expertise is particularly lacking e.g. AI, advanced analytics and data architecture. Where external partners/vendors deliver training, a qualification/certification system would help in understating quality and competence. Training on methodology and data processing needs to be implemented as well as on the characteristics of the data itself. Finally, a request to remember that industry could also be useful partners in the delivery of training.

- Communication

Communication was highlighted in the summary report as a key need but was re-emphasised in many of the stakeholder comments. The need for constant and proactive communication of the value of data sharing to patients and healthcare professionals was highlighted but also that alongside such discussions honesty was required around the potential risks of re-identification as well as of the rights of the patient as a data owner. In addition, the need to include the patient in all conversations was repeatedly mentioned; ensure the patient and healthcare professionals are included as a stakeholder when collaborative opportunities are identified, view patients as decision makers not simply as data generators, consult patients and healthcare professionals as partners during the development of any framework and finally involve patients in the communication of the awareness of data sharing rather than view them simply as needing to be convinced of its value.

- General comments

There was discussion about the proposed definition of Big Data. For example, it was suggested it be expanded to include structured and unstructured data, acknowledge in the definition the variety of data and note that in addition to revealing patterns, trends and association it potentially may bring automation and improved accuracy to decision-making. In the next phase of the BDTF, recommendations need to be prioritised to identify concrete and hands on ways to implement the ideas and recommendations in addition to identifying resources that will be made available.

9.2. Annex II: Taskforce membership

Name	Email	Agency	Subgroup	Subgroup	Member Since
Aldana Rosso	ALDR@dkma.dk	DKMA	Regulatory Acceptability		2017
Alison Cave	Alison.Cave@ema.europa.eu	EMA	Co-Chair	Policy, Training & Communications	2017
Antti Hyvarinen	Antti.Hyvarinen@fimea.fi	FIMEA	Analytics		2018
Armin Koch	Koch.Armin@mh-hannover.de	MHH	Regulatory Acceptability	Analytics	2019
Armin Ritzhaupt	Armin.Ritzhaupt@ema.europa.eu	EMA	Devices and in vitro diagnostics		2019
Astrid Schaefer	Astrid.Schaefer@bfarm.de	BFARM	Awaiting		2019
Aziz Diop	Aziz.DIOP@ansm.sante.fr	ANSM	Analytics		2019
Britta Ballhausen	britta.ballhausen@bvl.bund.de	BVL	Policy, Training & Communications		2019
Beatriz Sanchez	bsanchezd@aemps.es	AEMPS	Research Initiatives		2019
Cesar Hernandez Garcia	chernandezg@aemps.es	AEMPS	Research Initiatives		2017
Claes Enøe	CLEN@dkma.dk	DKMA	Research Initiatives		2019
Elizabeth Scanlan	Elizabeth.scanlan@ema.europa.eu	EMA	Policy, Training & Communications		2019
Fabrice Eroukhmanoff	Fabrice.Eroukhmanoff@legemiddelverket.no	NOMA	Research Initiatives		2019
Falk Ehmann	falk.ehmann@ema.europa.eu	EMA	Devices and in vitro diagnostics		2019
Florian Lasch	florian.lasch@ema.europa.eu	EMA	Regulatory Acceptability		2019

Name	Email	Agency	Subgroup	Subgroup	Member Since
Gianmario Candore	Gianmario.Candore@ema.europa.eu	EMA	Analytics	Research Initiatives	2018
Johannes Hendrikus Ovelgonne	h.ovelgonne@cbg-meb.nl	CBG-MEB	Devices and in vitro diagnostics		2017
Ivana Hayes	ivana.hayes@ema.europa.eu	EMA	Devices and in vitro diagnostics		2019
Jose Luis Alonso Lebrero	jalonsol@aemps.es	AEMPS	Research Initiatives	Policy, Training & Communications	2019
Jim Slattery	Jim.Slattery@ema.europa.eu	EMA	Analytics		2019
Katja Neubauer	Katja.Neubauer@ec.europa.eu	DG Santé	Data Quality and Standards		2018
Luis Correia Pinheiro	luis.pinheiro@ema.europa.eu	EMA	Analytics	Research Initiatives	2018
Marcel Maliepaard	m.maliepaard@cbg-meb.nl	CBG-MEB	Devices and in vitro diagnostics		2019
Marek Lehmann	Marek.Lehmann@ema.europa.eu	EMA	Data Quality and Standards		2019
Maria Kovacova	Maria.Kovacova@sukl.cz	SUKL	Analytics		2019
Marianne Van Heers	Marianne.VanHeers@ema.europa.eu	EMA	Policy, Training & Communications		2019
Anna Maria Gerdina Pasmooij	am.pasmooij@cbg-meb.nl	CBG-MEB	Devices and in vitro diagnostics		2017
Mark Goldammer	Mark.Goldammer@pei.de	PEI	Regulatory Acceptability		2018
Massimiliano Falcinelli	Massimiliano.Falcinelli@ema.europa.eu	EMA	Data Quality and Standards		2019
Nikolai Constantin Brun	NCBR@dkma.dk	DKMA	Co-Chair		2018
Nina Hein-Fuchs	Nina.Hein-Fuchs@pei.de	PEI	Analytics		2019

Name	Email	Agency	Subgroup	Subgroup	Member Since
Norbert Benda	Norbert.Benda@bfarm.de	BFARM	Awaiting		2019
Panagiotis Telonis	Panagiotis.Telonis@ema.europa.eu	EMA	Data Quality and Standards		2019
Paolo Alcini	Paolo.Alcini@ema.europa.eu	EMA	Data Quality and Standards	Analytics	2017
Peter Arlett	Peter.Arlett@ema.europa.eu	EMA	Co-Chair	Regulatory Acceptability	2019
Jens Piero Quartarolo	JEPQ@dkma.dk	DKMA	Research Initiatives		2019
Radim Tobolka	Radim.Tobolka@sukl.cz	SUKL	Awaiting		2019
Randi Munk-Jakobsen	RMJA@dkma.dk	DKMA	Policy, Training & Communications		2017
Renate Koenig	Renate.Koenig@pei.de	PEI	Data Quality and Standards		2017
Robin Seidel	Robin.Seidel@bfarm.de	BFARM	Devices and in vitro diagnostics		2019
Sara Rafael-Almeida	Sara.RAFAEL-ALMEIDA@ec.europa.eu	DG Santé	Regulatory Acceptability		2019
Thomas Sudhop	Thomas.Sudhop@bfarm.de	BFARM	Research Initiatives		2019
Vesa Kiviniemi	Vesa.Kiviniemi@fimea.fi	FIMEA	Regulatory Acceptability		2017
Yolanda Barbachano	Yolanda.Barbachano@mhra.gov.uk	MHRA	Regulatory Acceptability	Analytics	2019

9.3. Annex III: Blank Assessment fiche

https://www.ema.europa.eu/documents/template-form/hma-ema-joint-big-data-taskforce-phase-ii-report-annex-iii-blank-assessment-fiche_en.docx

9.4. Annex IV: Assessment fiches

https://www.ema.europa.eu/documents/template-form/hma-ema-joint-big-data-taskforce-phase-ii-report-annex-iv-assessment-fiches_en.docx

9.5. Annex V: Summary Table of Recommendations

BIG DATA TASK FORCE: DETAILED TABLE OF RECOMMENDATIONS

The Big Data Task Force vision is of a **strengthened regulatory system** that can efficiently **integrate data analysis** into its assessment processes to improve decision-making. This will be supported by **knowledge of data sources, their quality and their relevance for the European citizens**, **continual optimisation of data quality and analytical approaches** and promotion of a **secure and ethical data sharing culture**. **Training and external collaborations** will be key in order to build expertise.

Knowing when and how to have confidence in novel technologies and the evidence generated from Big Data will benefit public health by accelerating medicines development, improving treatment outcomes and facilitating earlier patient access to new treatments.

		Actions delivered by		
Focus	**Strategic Objective**	**Collaboration with external stakeholders**	**Consolidated action at level of EU regulatory network**	**Individual NCAs, EMA committees or working parties**
DATA ACCEPTABILITY	**TO ESTABLISH A FRAMEWORK WHICH DESCRIBES AND IMPROVES DATA QUALITY**	• Proactive communication of regulatory needs for data quality to funding bodies, data generators and academics. • Multi- Stakeholder workshops (including patients and healthcare professionals) to agree data quality metrics. • Engagement to promote adoption of the data quality framework (DQF), its implementation and transparency of the results. • Encourage and promote the use of ISO - IDMP for regulatory submissions of Big Data including real world data (RWD).	• Establish a DQF for regulatory use of Big Data sources to develop a common understanding of the strengths and limitations of Big Data sets. DQF must be model agnostic to allow a comparison across multiple different data sets. Initial funding is needed for a DQF for RWD but the methodology should be extended subsequently to other datasets which will have unique requirements. • Expansion and reinforcement of the existing qualification advice / opinion process and explore the merit and feasibility of an independent and renewable certification process for datasets and data collection methods.	• Define general data quality metrics for the DQF. • Contribute to guidance being developed or to be developed for the implementation of the new medical devices legislation and establish criteria to determine the accuracy, precision, reliability and comparability of device-based diagnostic tests and other in vitro diagnostics.

	TO DEFINE THE REPRESENTATIVENESS OF RELEVANT DATA SETS	• Promote the development of data sources in Member States where currently there are none. • Support and promote initiatives to access and link data across care settings.		• Develop guidelines to assess the representativeness and relevance of evidence derived from different populations and jurisdictions for European submissions.
	TO IMPROVE DATA DISCOVERABILITY	• Engagement with external data holders to enter information into the ENCePP Resources database. • Promote FAIR Principles and processes to sustainably increase data findability. • Utilise modern technology standards such as FHIR to allow data to be accessible, discoverable and exchangeable.	• Expand the scope and utility of the ENCePP Resources database to improve findability of RWD sets. This will require inclusion of more granular information on source, spectrum and quality (meta-data).	• Identify key data identifiers specific for regulatory decision-making to be included in ENCePP Resources database and FAIR Data Points (meta-data).
	TO SUPPORT ROBUST DECISION-MAKING	• Dissemination of guidance and best practice. • Seek alignment with guidance documents issued by other regulatory authorities.	• Building on the DQF, develop an iterative learning framework to define the evidential requirements for acceptability of new data types across the range of regulatory decisions. Considerations extend to data submitted as part of a regulatory procedure or generated externally via Big Data analyses Acceptability is determined in the context of the level of risk associated with each decision.	• Develop guidelines on study design and reporting including clarity on data transformations, database quality and choice, oversight and reporting. • Develop a reflections document on the acceptability of different data sources and analytics approaches for different regulatory use cases. • Expand the EU PAS registry for registration of protocols, amendments and results for regulatory submissions. • Gather learnings on the utility of RWE in drug development establish a dedicated RWE pilot programme and make the results public. • Build on the current CHMP work on description in the assessment report of biomarker and

				diagnostic testing, and in collaboration with medical-device authorities, establish thresholds for evidence required to include genomic information in the SmPC and label, to ensure consistent recommendations around the need for genomic testing in clinical practice.
PROCESS	**TO EFFICIENTLY INTEGRATE DATA ANALYSIS INTO REGULATORY DECISION-MAKING**	• Communication of requirements for raw data submissions to underpin regulatory decisions. • Establish contact points with experts at Notified Bodies and explore platforms for scientific discussion.	• Implement bi-modal architecture to provide space for data. exploration and experimentation • Investigate cloud technology for building Big Data and analytics infrastructure to deliver adequate transfer speed. • Initiate a proof of concept pilot on the assessment of Patient Level Data (PLD) from clinical trials and discuss findings with stakeholders. • Continue to expand the ability to access and analyse relevant datasets not submitted as part of an application (through collaboration with database custodians). • Strengthen coordination between the medicinal product and devices legal frameworks by developing strong links between medicines agencies, and authorities for devices and device notified bodies.	• With input across committees and relevant working parties, review learnings on experiences of different EMA committees in reviewing PLD and establish a pilot to test the utility and practical aspects of targeted PLD analysis in medicines assessment (initial focus on clinical trial data). • Create an EMA Expert Working Group on methods and analytics by combining the existing biostatistics, modelling and simulation, extrapolation and pharmacokinetics groups and enriching with real world data and advanced analytics expertise. • Enrich processes for scientific advice related to Big Data applications (including AI and proteomics questions) where there may be significant uncertainties which require raw data analysis for resolution.

DATA GOVERNANCE	**TO ENSURE A SECURE AND ETHICAL DATA SHARING CULTURE**	• Engagement with external initiatives on the implementation of new EU data protection regulations in creating guidelines and codes of conduct on secondary use of healthcare data on data protection and ethics. • Ensure needs of patient and healthcare professionals are embedded into data governance. • Support initiatives exploring novel technological solutions to facilitate data protection. • Support efforts to develop incentive models for data sharing by those managing, transforming and analysing data sets. • Encourage data sharing plans which should include FAIRification for regulatory required studies following authorisation.	• Monitor data protection legislation implementation and ensure targeted engagement of medicines regulators to elaborate our use cases for data. • Formation of an ethical advisory committee to advise on ethical aspects of regulatory use of Big Data. Patients and healthcare professionals should be represented. • Within multidisciplinary involvement and in line with HCPWP's and PCWP's work plans, initiate a pilot study to establish patient and healthcare professionals' views on aspects of data sharing including data protection and ethics.	• Identify regulatory use cases and concerns by EMA Committees and working groups on data protection and data ethics. • Centralised tracking of ethical and data sharing national use cases. • Set up an ethics advisory committee to develop a framework of data governance principles.
TRAINING AND EXPERTISE	**TO INCREASE NETWORK CAPACITY AND CAPABILITY**	• Establish collaboration with specialist academic centres to support training needs. • Host PhD/ MSc projects in data science to explore novel analytical solutions for regulatory use cases e.g. signal detection and validation as well as RWD to demonstrate effectiveness.	• Comprehensive skills analysis of the EU regulatory network • Development of high-level training curriculum on Big Data to meet immediate analytical needs. • In the light of the skills analysis and finalised Big Data strategy, EU-NTC in collaboration with EMA Committees to develop and implement a comprehensive long-term Big Data training strategy.	• Deliver training through the annual work plans and mandates of the committees and working parties.

	TO INCREASE EXTERNAL COLLABORATIONS TO PROVIDE EXPERTISE IN DATA SCIENCE	• Direct collaboration with international regulatory partners to share experiences and best practice and to collaborate on training workshops.	• HMA-EMA Joint Big Data Steering Committee to develop consistent external messaging on Big Data • Framework to enable knowledge exchange with academic centres of data science and key initiatives (e.g. IMI EHDEN, EC Digital Single market, AI, m-Health). • Short-term placements in academia and sites where data are collected to support the development of new skills. • Promote regulatory science centres as a vehicle to drive collaborative regulatory science projects.	• Propose regulatory research priorities for funders in Big Data area (ensuring alignment with the regulatory science strategy and public health and stakeholders' needs).

OPTIMISATION	**TO DRIVE THE CONTINUAL OPTIMISATION OF THE REGULATORY ASSESSMENT OF BIG DATA: DATA QUALITY AND REGUALTORY PROCESSES**	• Establish a stakeholder consultation platform (industry, government, academic, healthcare professionals and patients) to address Big Data related questions and processes and drive mutually beneficial pilots. • Seek international regulatory alignment of existing and new data standards. • Seek international regulatory alignment of methodological recommendations for Big Data studies. • Promote, support and drive international interoperability efforts e.g. common data models, common data elements, data structures, formats. • Promote international programmes to improve and harmonise the measurement of biomarkers. • Seek alignment on data and regulatory requirements with all actors in healthcare sector (HTA/payers). • Engagement with key initiatives (e.g. IMI, Horizon Europe, EC) and outreach to key academic initiatives (including relevant international ones).	• HMA-EMA Big Data steering committee for oversight of implementation, horizon scanning and to agree data science research priorities. • Development of a clear data standards strategy couple with a development framework and action plan to support more efficient regulatory review of IPD.	• Anchored in the new methods working party, develop expertise on RWD to drive regulatory strategy but also provide in depth scientific advice on RWE applications which should include the option of raw data analysis. • Enrich and rework the existing working party on pharmacogenomics, to cover proteomics and other omics technologies. • In line with the draft Regulatory Science Strategy develop e more agile and flexible guidance processes which allow faster updates in the context of a fast moving scientific landscape. • Targeted recruitment to fill expertise gaps.

	TO DRIVE THE CONTINUAL OPTIMISATION OF THE REGULATORY ASSESSMENT OF BIG DATA: ANALYTICAL APPROACHES	• Promote federated machine-learning (ML) approaches.	• Establish a federated network of advanced analytics centres, linked to EU regulatory agencies. • Establish ability (by the regulator or service provider) to validate studies submitted by companies. • Pilot: Develop new AI/ML approaches to detecting signals in EudraVigilance data. • Pilot: Launch an ADR Hackathon project. • Pilot: Proof of concept to link serious ADR data with genomic data, involving patients and HCPs.	• Development of a validation framework to ensure that new dynamic AI-based endpoints utilised in clinical trials correlate with clinical benefit and define the wider representativeness of AI algorithms (e.g. their consistent performance across patient populations).
COMMUNICATION	**DEVELOP AN OVERARCHING STRATEGY TO COMMUNICATE REGULATORY APPROACHES IN THE BIG DATA FIELD**	• Communicate key outputs from Big Data Task Force. • Proactive external communication to raise awareness of regulatory needs. • Support patient and healthcare professionals' awareness on the need of systematic collection of information on disease, treatments and outcomes.	• Building on the existing HMA-EMA network of communication professionals, identify a Big Data communication focal point in each agency. • Identify activities and key messages for co-ordinated external communication. • Establish Big Data communication materials. • Definition of metrics for reach and impact.	

BIG DATA INITIATIVES			
Initiative	***Collaborate and Influence***	***EU regulatory network***	***EMA committees/working parties/NCAs***
DATA **A**NALYSIS AND **R**EAL **W**ORLD **I**NTERROGATION **N**ETWORK: **DARWIN** ESTABLISH AN EU PLATFORM TO ACCESS AND ANALYSE HEALTHCARE DATA	• Engage proactively and promote initiatives to gain stakeholder support for an EU platform to access and analyse healthcare data (DARWIN). • Promote the need for sustainable and long-term funding for DARWIN via engagement with key European funding bodies. • Engage and outreach out to key academic/regulatory/national initiatives (including key international initiatives). • Liaise with key stakeholders, including industry, HTA and payer organisations on the business case for accessing EU healthcare data. • Ensure the business case includes the needs of different parties including patients, HTA and payers, national authorities, EU health agencies and the European Commission (in line with the vision for an EU health data space).	• Develop a clear business case which includes the delivery and sustainability model for DARWIN. • Network should be scalable and allow for differing speeds of adoption and implementation. • Establish an analytical system to support real time analytics in DARWIN).	• Develop regulatory use cases across EMA committees and working parties to inform thinking and ensure that DARWIN delivers for regulatory needs.

9.6. Annex VI: Biostatistics Working Party positions on Patient level data assessment

The question of whether patient level data[29] should be assessed as part of the authorisation procedure was considered by EMA management board in December 2014. At the time management board considered a deeper reflection was required: while the concept was largely accepted, a clarification of the objectives, the development of criteria which would trigger a patient level analysis and an examination of resource implications was requested.

Five years on, the data landscape has changed and even more change is anticipated in the next 5 years. Hence, as data availability increases and data from more sources are integrated into applications, it is the view of the BDTF that our current approach will become increasingly limiting and will potentially impact on the robustness of our assessments. Moreover, if the data is such that pre-specified, standardised analyses are not possible, simply requesting a re-analysis of the data by the company when uncertainties arise may not be sufficient. Other international regulatory agencies, including the US FDA and Japan PMDA, already receive PLD as part of regulatory submissions and use it to support their assessment of the marketing authorisation application dossier. Both agencies mandate CDISC standards for datasets and associated metadata for marketing authorisation applications. If we are to move towards a system where we can efficient integrate data analysis into our decision-making, we need a radical change in approach and processes.

Importantly the legal framework for assessment of PLD is provided by Annex I of Directive 2001/83/EC which lists the particulars and documents required for a marketing authorisation application '*all information that is relevant to the evaluation of the medicinal product concerned, shall be included in the application, whether favourable or unfavourable.'* In the context of clinical study reports and their contents, the Annex states that the clinical particulars provided with an application '*must enable a sufficiently well-founded and scientifically valid opinion*' (5.2 a) and must contain '*sufficient detail to allow an objective judgement to be made*' (5.2 d).

In processing PLD, and to the extent that the data are not fully and irreversibly anonymised, the Agency is bound to comply with the provisions set in the Regulation (EC) No. 45/2001 on the protection of personal data. Equally we must ensure that any development align with the implementation of the clinical trial Regulation (EU) No. 536/2014 and the development of the EU clinical trial portal.

In order to ensure the concerns of management board from 2014 are adequately addressed the Biostatistical Working Party developed a position paper in 2018 (Annex VII) with two key recommendations which are supported by the BDTF (Fiche #1):

- **Formation of a cross committee working group**: this group would examine the practical aspects of PLD analysis. More specifically it would define the particular circumstances in which PLD assessment would add value including the expected impact of such analysis on the timescales of the relevant procedures and the resourcing (human and financial) needs. In addition, the working group would establish the requirements for technical infrastructure, data standards and tools needed for an initial use of PLD and include a technical development path adequate to the foreseen needs. Fiche #1 (attached file) sets out the case for consistent data standards which as far as possible should be minimised and harmonised internationally (core recommendation of Phase 1

[29] IPD is defined as data, including imaging data, at an individual patient level which is directly assessable in terms of re-analysis or additional analyses.

Summary report). For clinical trial standards the Clinical Data Interchange Standards Consortium (CDISC) standards, already mandated by FDA and PDMA, would be the obvious choice and would align with the principles of minimisation and harmonisation. These standards are developed and maintained by CDISC and cover both standards to aid data collection at clinical investigation sites and standards to structure and transmit the standards.

- **A Proof of Concept pilot**: To inform the thinking of the working group and specifically its estimation of human resourcing and technological needs a proof of concept pilot is suggested to examine the scenarios in which PLD assessments might add value. PLD from ten marketing authorisations would be requested: possible scenarios include:
  - where the pivotal trials showed a small magnitude of effect which is of clear borderline clinical value;
  - where the pivotal trials were single arm trials especially if the use of historical controls is proposed;
  - where the pivotal trials included new active substance or new ATMP
  - for applications involving real world data to support effectiveness claims;
  - for small populations e.g. in paediatric trials;
  - where another regulator's conclusion is different to that of the EMA which may have arisen as a result of their analysis of PLD;
  - concern around fraudulent data.

Hence the BDTF fully supports the proposition about the potential added value and benefits of PLD use for the evaluation of benefit-risk of human medicines, upon request from EMA Scientific Committees, and within the boundaries set by the current applicable legislation. This recommendation sits predominantly at the level of the EU regulatory network.

9.7. Annex VII: Resources

https://www.ema.europa.eu/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-annex-vii-resources_en.xlsx

9.8. Annex VIII: DARWIN business case

Business case for an EU platform for accessing and analysing healthcare data (**D**ata **A**nalysis and **R**eal **W**orld **I**nterrogation **N**etwork **(DARWIN)**

Purpose of the paper

- To set out the business case for a platform for accessing and analysing EU health data with an initial focus on electronic health records (EHRs).
- The central premise of the business case is that access and analysis of EU health data will support early access to medicines thereby fulfilling the unmet medical needs of EU citizens, and will support a learning healthcare system for marketed products enabling safe and effective use of medicines.

Opportunities from analysis of healthcare data

- There is a dramatic increase in the digital capture of healthcare data providing access to unprecedented amount of information from EHRs, claims data, registries, lab data, images, and genomics data.
- Access and analysis of these data can inform regulatory decision-making throughout the product lifecycle:
  - Support product development (e.g. scientific advice, PRIME) informing on unmet medical needs, historical controls, demographics of the population to be treated, choice of dataset for longer-term follow up.
  - Support authorisation of new medicines – for both initial authorisation and line extension (complementing randomised clinical trials). Regulators will be able to contextualise the data submitted by industry and when appropriate validate study findings.
  - Monitor the performance of medicines on the market (effectiveness and safety) using real world data (RWD).
  - If knowledge is fed into decision-making, enables a 'learning healthcare system'.
- With major use of real-world data and big data by the pharmaceutical industry, regulators need to have the ability to perform targeted validation of claims through independent analysis.
- Additional benefits will come as EU partners and stakeholders participate and access the platform, including: European Commission for policy, impact and monitoring (e.g. climate and health), EU health agencies, National governments including HTA bodies and payers.

EU landscape for healthcare data

- The EU has a rich and diverse data landscape:
  - With many healthcare systems being state funded, long-term follow up of patients is possible (in contrast to the U.S. situation where patients frequently move between insurance schemes).
  - Diversity also brings challenges in terms of language, data structure and approaches to governance.

- An analysis of the accessibility of different EU health data sets suggests at least eight different access frameworks ranging from no access to commercial sale of pseudo-anonymised data.
- EU data protection legislation requires navigation but is compatible with the secondary use of healthcare data for justified public health and research purposes (although work is required in this area to provide clarity to stakeholders on how to comply with legislation).
- The EU has a vigorous academic environment with the development of innovative approaches to extract, standardise and combine health care data available in different formats, but with funding constraints allowing only short-term projects (EU funding to date has been entirely project based with no sustainable funding mechanism put in place for an EU RWD platform).
- Stakeholders and initiatives are aligning:
  - The European Commission is currently coordinating the EU eHealth network to support data access and sharing and there are discussions on an EU Health Data Network. This forum of Member State health ministries can be leveraged to gain support for an EU health data platform.
  - The draft EMA Regulatory Science Strategy includes prominent actions on healthcare data accesses (big data and RWD). The consultation response has confirmed use of RWD as being in the top three priorities for stakeholders.
  - The EMA-HMA Big Data Task Force has as its principle recommendation an EU platform to access and analyse healthcare data.
  - Work to deliver health data access can be underpinned by the EU regulatory network Strategy to 2025.
  - The Council and the Commission are both calling for a framework for post-authorisation vaccine studies to ensure robust evidence on vaccine safety and effectiveness, thereby combatting vaccine hesitancy (this proposal is complementary and could be combined).
  - Sustainable funding can potentially be obtained through the revision of EMA fees regulation.
  - The new Commission is working on its multi-annual funding and 'digital', and 'data' feature prominently suggesting that EU funding to set up a healthcare data platform could be made available.

Benchmarking with international regulators

- The U.S. Food and Drug Administration (FDA) has invested close to a billion U.S. Dollars on RWD over the past ten-years (with the Sentinel system at its core but also including various complementary data access and analysis approaches). This was based on explicit legal basis and funding including the 21st century cures act. In October 2019 FDA has announced a new $220 million 5-year contract with Harvard Pilgrim for the Sentinel program.
- In Canada, Health Canada and the health research directorate have established CNODES a federated network of healthcare databases in their provinces that runs common protocol studies on RWD. The annual operational funding of the Canadian model is approximately Euro 5 Million.
- In Japan MHLW and PMDA have collaborated to establish a hospital data network system to enable analysis of RWD.
- Other stringent regulators around the world have and continue to develop sustainable platforms to access and analyse healthcare data with a primary focus on EHRs and insurance claims data.

Current experience and access for EU regulatory network

- A minority of National Competent Authorities (NCAs) including but not limited to those in the UK, Denmark, France, and Spain can access and analyse national RWD. A number of additional NCAs are actively developing competence and data access. A recent review of electronic health care databases in Europe identified 34 databases in 13 EU countries, with variable level of access [3]. Access to health care data is however gradually enabled and expanded in several countries.
- EMA has purchased access to pseudo anonymised electronic health records from the UK, France and Germany and additionally contracts are in place with academic consortia allowing RWD studies from a larger number of datasets from across the EU (36 datasets from 10 MSs). From 2013 to 2019, EMA has conducted 74 studies with in-house electronic health record databases and commissioned 18 external studies. These have mainly been in the area of drug safety and have supported decisions by the PRAC and CHMP.
- EU projects have laid strong foundations for a sustainable platform and examples of what has been delivered are provided at 'DARWIN Background 2'.

Who could benefit from improved access to and analysis of EU health data

- The principle use cases and benefits outlined in this paper relate to the regulation of medicines and access to an EU heath data platform for EMA and NCAs is therefore core. However, through multi-stakeholder collaboration in the establishment and operation of the platform, additional use cases and benefits will be delivered and a cooperative model should reduce costs, increase the accessibility of data and build political support. As such access could be foreseen for:
  - EU and national regulators;
  - HTA bodies;
  - Payers;
  - Health ministries;
  - European Commission;
  - EU health agencies;
  - And possibly accredited EU patient and healthcare professional associations, and academia.
- It is possible that more data sets will be made available if access is restricted to public bodies with an explicit health mandate. Therefore, there should be a debate on access for the pharmaceutical industry and if access is granted further debate on the terms and governance for such access.

Design principles to improve access to and analysis of EU health data

- We consider that one monolithic system is not optimal. Despite the Sentinel system, U.S. experience shows that bespoke solutions are needed for specific use cases and product types. Furthermore, the healthcare data landscape is evolving very quickly and our approach should therefore be agile.

- The following design principles for the platform are suggested:
  - Start with access to EHRs;
  - Access hospital (specialist use) as well as GP records;
  - Have EHRs in house (or get direct remote access) for rapid analytics to support immediate committee decision-making (pilots with PRAC and CHMP planned for 2020);
  - Use a federated 'common protocol' model to access data from a large number of MSs with different data access patterns and for more complex studies including when causal inference is required;
  - Establish a data quality framework to inform regulatory decisions on use of particular datasets and to support regulatory acceptability;
  - Establish governance rules and processes including access rights and data protection arrangements;
  - Leverage expertise from academia and NCAs;
  - Establish a technology platform for data exchange, management and analysis;
  - Collaborate with stakeholders to increase data access and reduce costs;
  - Establish funding for the set up and maintenance phase (sustainable funding).

Funding

- Two phases are foreseen: set-up and operation:
  - Set-up phase will need project funding at 30 - 50 Million Euros (ROM). This will deliver a technology platform and establishment of governance, quality frameworks and partnership agreements. It is proposed that this comes from the EU budget e.g. DG RTD or DG CONECT.
  - Maintenance phase will need 10 – 20 Million Euros annually (ROM). This will cover maintenance and evolution of the technical platform, maintenance of the governance framework, delivery of core routine data analyses and a small number of regulator requested studies. One potential source is that this comes from EMA fees (although this would require a change to the fees regulation). Additional funding for specific studies could come from different stakeholders (based on their needs for study results).

Benefits to stakeholders

- Benefits include:
  - EU and national regulators – data analyses support better quality decisions on the development, authorisation and supervision of medicines' benefit risk;
  - HTA bodies and payers - data analyses support better quality decisions on cost-effectiveness, optimising healthcare spending and targeting therapies at patients most likely to benefit;
  - Health ministries – data analyses support health policy development including design and delivery of healthcare systems;

- European Commission – data analyses support health policy and legislative development and monitoring of implementation. The platform will support emerging needs including monitoring of climate change and health;
- EU health agencies – use cases specific for EFSA, ECDC, ECHA, JRC;
- EU patients – faster access to innovative medicines and optimisation of safe and effective use.

Conclusion

- The EU regulatory network is falling behind international partners in the area of accessing and analysing RWD. However, we still have an opportunity to accelerate product development and optimise use of products on the market by enhancing decision-making using health care data.
- To realise this potential will require a major initiative including support from across our stakeholders.
- The time is now with EMA, EU Network and EC initiatives aligning with the needs and desires of stakeholders to leverage data for better healthcare.

DARWIN Background 1: Examples of national initiatives in the EU:

- In Finland new legislation entered into force on 1 May 2019 which opened up the national social and healthcare registers linked to patient systems in primary care, specialist healthcare and social services to allow their secondary use by external requesters. The 'Findata' platform will be a one-stop shop open to requests from January 2020 with provision of data within 60 days after approval, instead of 2 to 3 years previously.
- In France, the law expanded since April 2017 the access to the data of the national health care system on nearly 67 million people comprising outpatient claims data, hospital summaries and data from the death registry. This database is being further developed to create a health data hub aiming to link other data sources.
- In Estonia, the Estonian Biobank is a prospective longitudinal database covering 5% of the adult population (more than 50,000 participants) and including results of DNA, plasma and cell samples with linkage to electronic health records and patient registries. The Estonian Human Genes Research Act enforced since January 2001 opened the Biobank for research based on clear access rules and a broad informed consent.
- In the UK, Health Data Research UK (HDRUK): had an initial investment of £120M, now £37M per year (2,700 people recruited) to enable innovation based on access to healthcare data.
- In Spain, AEMPS created and funded in 2001 the BIFAP database (Base de Datos para la Investigación Farmacoepidemiológica en Atencion Primaria), a computerised database of medical records of primary care representing 16% of the Spanish population and including data on around 9 million patients.

DARWIN Background 2: EU projects have laid strong foundations for a sustainable platform and examples of what has been delivered are provided below

- PROTECT was one of the first private-partnership projects conducted under the IMI framework; it tested and applied an EU-wide common protocol model for multi-database studies and examined conditions of success for such networking model.
- The European Network for Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Database of Research Resources may facilitate such networking by allowing the identification of centres and data sets by country and type of research.
- Several research projects have also developed (and are still developing), the tools and infrastructure for the mapping and utilisation of European data sources in a common data model within a federated data network. By supporting data standardisation and implementation of a common quality framework, these projects will facilitate performance of high quality and reproducible studies across Europe. Data can be extracted from local databases using a study-specific, database-tailored extraction into a simple common data model (CDM). The resulting data can be transmitted to a central data warehouse as patient-level data or aggregated data for further analysis. Examples of research networks that used this approach by employing a study-specific CDM are EU-ADR, SOS, ARITMO, SAFEGUARD, GRIP, EMIF, EUROmediCAT and ADVANCE.
- Several organisations are also applying a generalised CDM (conversion of the totality of the database) similarly to the US Sentinel and OHDSI projects. The main advantage of a general CDM is that it can be used for virtually any study involving that database. OHDSI is based on the Observational Medical Outcomes Partnership (OMOP) CDM and is partnering with IMI-funded EHDEN project that aims to build a large-scale, federated network of data sources standardised to a common data model in Europe.