13 February 2019
EMA/105321/2019

HMA-EMA Joint Big Data Taskforce

Summary report

See websites for contact details

Heads of Medicines Agencies www.hma.eu
European Medicines Agency www.ema.europa.eu

The European Medicines Agency is
an agency of the European Union

Table of contents

1. Abstract

The increasing volume and complexity of data now being captured across multiple settings and devices offers opportunities for medicines regulation in terms of a better understanding of diseases, medicines and the performance of products in the healthcare system. However the acceptability of these data in the regulatory context requires an understanding of their provenance and quality in addition to the validity of new approaches and methods for processing and analysing these data e.g. using algorithms and machine learning. There is little doubt that drug development will increasingly utilise these data and hence a regulatory strategy is required to determine when and how in the product life cycle evidence derived from such data may be acceptable for regulatory decision making. The HMA–EMA Joint Big Data taskforce was formed to describe the big data landscape from a regulatory perspective in order to ensure the EU regulatory system has the capability and capacity to guide, analyse and interpret these data.

Although the term of big data is widely utilised there is not one commonly accepted definition. Any definition should encompass not only the concept that big data is diverse, heterogeneous and large and incorporates multiple data types and forms but also should refer to the complexity and challenges of integrating the data to enable a combined analysis. Considering the regulatory context, the taskforce defined big data as *'extremely large datasets which may be complex, multi-dimensional, unstructured and heterogeneous, which are accumulating rapidly and which may be analysed computationally to reveal patterns, trends, and associations. In general big data sets require advanced or specialised methods to provide an answer within reliable constraints'.* Such methods of data analysis i.e. models and algorithms are the subject of the growing field of data science, which combines methods from various disciplines and is critical for harnessing the opportunities from big data.

The taskforce committed to consider all data relevant to regulatory decision making and hence formed six subgroups to assess the data landscape in genomics, bioanalytical 'omics (with a focus on proteomics), clinical trials, observational data, spontaneous ADR data and social media and m-health data. A cross cutting data processing and analytics group was formed in early 2018 and is due to deliver a report in the first half of 2019.

The current report is an overarching summary of the reports of the six subgroups with recommendations setting out a number of common steps along the road to regulatory acceptability of big data. The recommendations start with data standardisation in order to define and, where possible, improve data quality and progress to actions to promote data sharing, access and enable robust big data processing and analysis. These steps are not necessarily sequential, are often interdependent and will be iterative requiring a constant engagement with external stakeholders to articulate regulatory needs as the field further develops. Additionally across all big data domains there is a need to ensure the European regulatory network develops and maintains sufficient expertise to guide, interpret and critically assess big data.

Specific recommendations from each of the subgroups are tabulated and prioritised as an annex to this summary report. While data linkage was a common theme, the need to link genomics data with clinical outcomes was emphasised; additional challenges included achieving timely update of clinically relevant genomic information in medicinal product information and defining performance standards for companion diagnostics. For proteomics the key need is for technical and analytical standards and, as utilisation increases, the provision of timely regulatory guidance to promote the development of validated proteomic biomarkers. The vast majority of clinical trials are never submitted as part of a regulatory submission and standardisation activities are critical to increase data interoperability and facilitate data sharing. For observational data, often referred to as real world data, which includes data from electronic health records, there is an urgent need to enable timely access to data representative

of the European population but also to increase the capacity to analyse and integrate such evidence into decision making across the product life cycle. Databases of spontaneous adverse drug reports (ADRs) already meet mandatory standards but the increasing volume of reports require new analytical approaches to efficiently analyse and link ADRs and in particular exploit data contained in unstructured case narratives. Opportunities exist for m health/wearables to deliver new outcome measures and record lifestyle factors currently not possible to reliably capture via other means, but standards need to be defined if this data is to be used for regulatory submissions. Finally, the most natural application of social media data is for pharmacovigilance and signal detection but further research is required to identify the specific areas where such data will add most value.

It is clear that big data sets vary in structure and quality and hence some will be more immediately relevant for regulatory decision making. Thus a prioritisation and focusing of activities is required, which should build on ongoing actions and where possible link to existing structures, to deliver outputs in a timely manner. Notably the current report sets out ‘what‘ needs to be addressed but ‘the which‘, ‘the how‘ and ‘the when‘ will need further work and as such the mandate for the HMA-EMA Joint Big Data Task Force has been extended. The scope of work is large and urgent but we must not step away from these challenges; in some areas we are already moving in the right direction but not in a consistent and consolidated way and we therefore need to guard against reverting to the status quo. Rather, when challenged with new scientific and technological possibilities we should engage in order to ensure we have the capability and capacity to guide, analyse, interpret and provide benefit for patients from the data generated. In this way we will improve our decision making and enhance our evidentiary standards.

2. Introduction

Medicines regulation is facing multiple challenges. The unparalleled pace of change in the scientific landscape is driving a paradigm shift in drug development, challenging regulatory agencies to look beyond conventional sources of evidence to support decision making across the entire product life cycle. These new sources of evidence, often collectively termed big data, offer opportunities to improve decision making but also bring uncertainties around the quality of the data and hence the evidence generated.

There are many potential opportunities: data from real world populations will provide the possibility to evaluate safety and effectiveness in normal clinical practice. Capturing comprehensive healthcare data across care settings will not only better inform clinical trial design but also deliver better ways to monitor benefit risk over the long term. Big data strategies[1] may deliver insights into physiological and pathophysiological processes at the level of the organ, the cell[2] and even at the subcellular level while the development of novel pharmacological models[3] could improve the efficiency and cost effectiveness of model-informed drug discovery and development. With such approaches we may be able to better characterise and predict adverse drug reactions but also identify biomarkers of disease and drug responsiveness. Lastly and perhaps most obviously, advances in genomic sequencing techniques and gene therapy are offering innovative, targeted treatment possibilities which is supporting development of treatments in orphan diseases, and, enabling a move from broad, one size-fits-all indications towards narrower, more precise indications. This is driving an almost explosive increase in the development of *in vitro* diagnostic tools and methods necessary to characterise and stratify disease to better identify the target populations.

In parallel to the rise of ‘omics technologies and complex modelling, is the development of wearables and the growing number of m-health apps, driving the movement of the “quantified self” but also enabling virtual clinical trials whereby data collection via mobile phones, tablets and other telemedicine services enable patients to participate in clinical trials from home, regardless of geographical location. Such approaches may bring novel insight and offer opportunities to capture a holistic picture of the patient including previously uncaptured lifestyle factors but may also deliver unstructured and fragmented data and thus decrease trust in the evidence that is generated, especially in the context of regulatory decision making.

There is little doubt that the development of future treatments will utilise such big data. Data may reach regulatory authorities either as supportive data, in the margins of more traditional analysed structured data, or may underpin the submission as a whole. It is thus essential that the regulatory network understands its presence and the robustness by which it was generated in order to make a competent evaluation of the submission as a whole.

This challenge is significant: our regulatory framework for authorisation is based on the assessment of well-controlled, randomised, high-quality data of known provenance with the aim of identifying a relevant clinical difference between the medicine in question and a given control in order to provide unbiased estimates of efficacy and safety. In contrast, big data offers evidence which may be derived from unstructured, heterogeneous and unvalidated data of unknown provenance and unknowns around potential bias with additional uncertainties of accuracy and precision. Moreover not all datasets are the same; there is variable quality and standardisation, data is generated under different scenarios and for different purposes which rarely includes medicines regulation and ownership resides with multiple stakeholders many of which have no need to engage with the regulatory system. Influencing the data

[1] https://fair-dom.org/partners/virtual-liver-network-vln/
[2] http://plateletweb.bioapps.biozentrum.uni-wuerzburg.de/plateletweb.php
[3] http://www.ddmore.eu/

landscape to meet regulatory needs is thus complex. As a regulatory network, we must prepare for and understand the change in data generation and knowledge management. This requires harmonisation of business processes, IT systems and a strategy as to how and when in the product life cycle the utilisation of such data can bring value to our assessment of new therapies.

The HMA-EMA Joint Big Data taskforce was formed upon this background with a mandate 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.

This report summarises the main outputs of the task force[4].

3. HMA-EMA Joint Big Data taskforce Membership

The taskforce is composed of representatives from 14 National Competent Authorities (NCAs) plus EMA representation and until February 2018 was chaired by Thomas Senderovitz, Danish Medicines Agency and Alison Cave, EMA. From February 2018, Nikolai Brun of the Danish Medicines Agency replaced Thomas Senderovitz as Chair of the taskforce. The full list of taskforce members can be found at Annex 1.

4. Definition of Big Data

Although the term big data is widely utilised there is not one commonly accepted definition of big data. Any definition should encompass not only the concept that big data is diverse, heterogeneous and large and incorporates multiple data types and forms but also should refer to the complexity and challenges of integrating the data to enable a combined analysis. Hence the taskforce defined big data as *'extremely large datasets which may be complex, multi-dimensional, unstructured and heterogeneous, which are accumulating rapidly and which may be analysed computationally to reveal patterns, trends, and associations. In general, big data sets require advanced or specialised methods to provide an answer within reliable constraints'.* Thus, a single dataset does not strictly meet the definition of big data e.g. a single clinical trial dataset or electronic health records from a single supplier. However, when pooled with other datasets of a similar type, or linked to other datasets of different types through data sharing initiatives, the datasets become sufficiently large or the difficulties in pooling, linking and analysing are sufficiently complex, for the data to assume the characteristics of big data

As a result, the taskforce committed to consider all big datasets regarded as relevant to regulatory decision making, including those datasets which if pooled and linked would be regarded as big data. The following subgroups were formed:

- Clinical trial and Imaging subgroup;

[4] The scope of work included in this report did not include a consideration of the legal requirements to protect patient privacy when sharing data. As such the recommendations outline in this report will need to consider at all times data protection and meet the requirements of the of EU data protection legislation, in particular Regulation (EU) 2016/679 (the General Data Protection Regulation) or Regulation (EU) 2018/1725, as applicable. This will be incorporated into the next phase of the work.

- Observational data subgroup;
- Spontaneous adverse drug reports subgroup;
- Social media and m-health subgroup;
- Genomics subgroup;
- Bioanalytical ‘omics subgroup;
- Data analytics subgroup[5].

5. Scope of the taskforce

The work of the taskforce focussed only on data related to humans and was divided into 4 main workstreams which are described below.

5.1. Workstream 1

Each subgroup was asked to characterise its dataset which should include where possible consideration of: (i) data structure (ii) data provenance; (iii) data quality; (iv) data heterogeneity (v) speed of change and/or rate of accumulation; (vi) completeness and opportunities to capture data; (vii) data standards and terminologies; (viii) data accessibility; (ix) analytical methodologies; (x) uncertainties or unknowns which require further exploration; (xi) specific regulatory challenges.

5.2. Workstream 2

To inform thinking, the range of regulatory decisions was mapped across the product life cycle (Annex II) and each subgroup was asked to consider specific areas of usability and applicability where big data may add value. In considering gaps and opportunities subgroups considered:

- Whether the status quo is satisfactory;
- Whether there are areas where decisions are made on inadequate data and if so where;
- Where technological advances may create uncertainties;
- Where technological advances will offer opportunities to resolve uncertainties and improve our decision-making.

5.3. Workstream 3

In order to ascertain the current situation across the European regulatory network with regard to the available expertise and competences for the analysis and interpretation of Big Data, a survey was launched to define:

- the current state of expertise and experience of NCAs regarding Big Data;
- plans to increase resources where expertise is scarce;
- the current challenges identified by NCAs and;
- the expectations of and future challenges anticipated by NCAs.

[5] The work of the data analytics group is ongoing. The group will report by end of Q1 2019.

In addition, an e-survey was launched addressed to pharmaceutical companies that sought to understand the current experience, key challenges, applicability and added value of big data in the context of the life cycle of a product.

5.4. Workstream 4

The final deliverable of the taskforce is a list of recommendations and a Big Data Roadmap.

6. Outputs of the taskforce

6.1. Output from Workstream 1 and 2

Each subgroup performed an extensive analysis of the data landscape relevant to the data category considered by the subgroup and considered specific areas of usability and applicability across the product life cycle. Each report delievered a specific set of recommendations that support the core recommendations described in workstream 4. These recommendations are provided at Annex III.

6.2. Output from Workstream 3

Two surveys were launched by the taskforce. The first survey for completion by the NCAs was to ascertain the current situation across the European regulatory network with regard to the available expertise and competences in the analysis and interpretation of big data. The second survey was intended to ascertain the current landscape in terms of application of big data by industry across the drug development pathway. Electronic versions of both surveys are attached in Annex IV.

6.2.1. Synopsis of the results of the survey of the National Competent Authorities

A full graphical representation of the results from the survey, which was completed by 24 out of a possible 33 NCAs, is provided in Annex V. The results of the survey demonstrate that there is currently very limited expertise in big data analytics at national level but this partly reflects a belief that such expertise is not needed at the current time especially given there has been only limited demand from industry through national scientific advice procedures[6]. In addition, 8 of 24 NCAs reported no in house expertise in biostatistics, which is a key analytical need even at the current time. This does not include the statistical and analytical expertise available within the EMA. The majority of NCAs believe an increase in expertise in this area will be required within a 5-year timeframe but few had concrete plans in place; where timescales were specified they were mainly for action within 3 years. It important to note that the term big data may have been viewed differently by different NCAs and interpretation of the terms around specific expertise may have varied. The taskforce did not re-contact NCAs to ask for clarification with regard to any responses.

More than half of NCAs currently have direct access to external big data sets (mainly RWD / observational data, ADR data and clinical trial data) and relevant in-house systems / tools to meet current analytical needs (mostly SAS, R). Determining future needs will be a key deliverable of the data analytics subgroup[7].

[6] Only 3 NCAs report receiving requests relating to the applicability or analysis of big data sets through the scientific advice process.
[7] Report and recommendations will be delivered in 2019.

Few external collaborations with academic institutions (only 6/17 NCAs) were reported[8] and it seems clear that maintaining sufficient expertise within the regulatory network will be an increasing challenge. Establishing strong collaborative links with academic institutions will be necessary to support training needs and strengthen the capacity of the European regulatory network.

In line with the limited experience through scientific advice, few NCAs reported receiving evidence derived from big data sets within regulatory procedures. The greatest experience was in support of pharmacovigilance needs during post authorisation procedures including PASS, PAES, RMPS, PSURS and referrals and, where specified, related mostly to the use of observational datasets e.g. registries, claims and EHRs. Signal detection, validation and assessment were identified as the regulatory areas where big data is likely to be applied first for which natural language processing (NLP) would be a key advance to interrogate case narratives. Other key areas including clinical trial data sharing and the validation of benefit-risk in high risk populations often excluded from clinical trials.

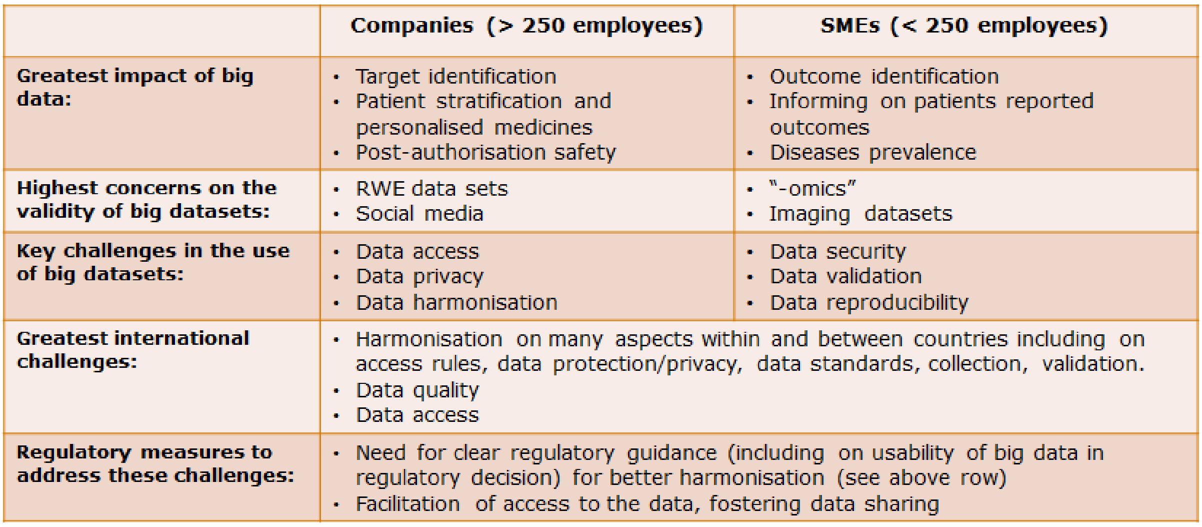
Data quality is considered the biggest challenge for the use of big data for regulatory decision-making followed by data harmonisation and integration across Europe and the ultimate validation of the derived evidence. In line with the above discussion, the need to increase expertise and capacity within the regulatory network was re-iterated.

6.2.2. Synopsis of the results of the survey of industry

The response rate of the survey by industry was lower than expected (37); nevertheless the profile of the responses were almost equally split between large pharma (>250 employees) and small to medium enterprises (SMEs; 10 to 250 employees) which allowed a comparison of views and priorities. Interesting discrepancies in areas of focus were revealed which are summarised in Table 1, which suggests different areas of business focus for these organisations. For example, large pharma highlighted personalised medicine, understanding current clinical care and informing clinical trial design as areas where big data would have the greatest potential impact while SMEs highlighted patient reported outcomes, outcome identification and signal validation as the most important areas. This may be reflective of a focus of SMEs on technology-related opportunities such as m-health apps and wearables and a focus on new treatment options for unmet need in rare diseases.

Challenges highlighted by industry participants mirrored those of NCAs, with data access, data integration, data validation and data reproducibility all highlighted as key concerns. In addition, data security and data protection were key concerns and these issues are included as reinforcing actions underpinning the core recommendations of data sharing and data linkage and highlighted within the recommendations of the clinical trial subgroup.

[8] The question focussed only on external academic collaborations and did not include collaborations across NCAs such as within Committees or working parties.



	Companies (> 250 employees)	SMEs (< 250 employees)
Greatest impact of big data:	• Target identification • Patient stratification and personalised medicines • Post-authorisation safety	• Outcome identification • Informing on patients reported outcomes • Diseases prevalence
Highest concerns on the validity of big datasets:	• RWE data sets • Social media	• "-omics" • Imaging datasets
Key challenges in the use of big datasets:	• Data access • Data privacy • Data harmonisation	• Data security • Data validation • Data reproducibility
Greatest international challenges:	• Harmonisation on many aspects within and between countries including on access rules, data protection/privacy, data standards, collection, validation. • Data quality • Data access	
Regulatory measures to address these challenges:	• Need for clear regulatory guidance (including on usability of big data in regulatory decision) for better harmonisation (see above row) • Facilitation of access to the data, fostering data sharing	

Table 1: Synopsis of the results from the industry survey

The survey specifically asked 'What measures could the regulatory network introduce to address the highlighted challenges?'. Specific actions noted by both large pharma and SMES were rules/regulatory guidelines /information in addition to provision of datasets. The latter is not usually within the remit of regulatory agencies although proactive actions such as Policy 0070[9] demonstrate regulatory support of data sharing principles and the taskforce recommendations list a number of key actions necessary to promote a data sharing culture among data owners. The last questions focused on international challenges and the perennial issues of data quality and access were raised in addition to the need to harmonise data both within and between countries. A key enabler for harmonisation is data standardisation and the recommendations highlight a number of actions including open source file formats and standards. A key additional comment that is strongly supported by the taskforce is the need for collaboration across all stakeholders from regulators to payers, HTA bodies, patients, academia and industry.

6.3. *Output from Workstream 4: list of recommendations*

New emerging data sources offer opportunities but also bring uncertainties around the evidence generated. As the vast majority of the data sources considered are not generated for regulatory approval, strategies are needed to better understand data quality in order to articulate where, when and how such evidence may be acceptable for regulatory submissions and subsequent monitoring of medicines. A consolidated table of the subgroup recommendations can be found at Annex III. From these recommendations a number of core recommendations emerged which focused on 9 key areas: (i) data standardisation, (ii) data quality, (iii) data sharing and access, (iv) data linkage, (v) data analytics, (vi) regulatory acceptability of big data analyses, (vii) medical devices/in vitro diagnostics regulation, (viii) skills and knowledge across the regulatory network and (ix) communication and engagement. A summary of these areas is provided below with each core recommendation accompanied by a number of reinforcing actions for which ownership is assigned. Where concerted European or global action is required a common ownership is assigned, where centralised regulatory oversight is the key requirement EMA ownership is assigned, where action must be driven by NCAs HMA ownership is assigned, with combinations of the above groups or other bodies included where appropriate.

[9] European Medicines Agency policy on the publication of clinical data for medicinal products for human use.

6.3.1. Data standardisation

Core Recommendation

Promote use of global, harmonised and comprehensive standards to facilitate interoperability of data

(supported by subgroup recommendations # 1, 6, 8, 9, 11, 12, 16, 32, 34, 44, 45)

- Minimise the number of standards; strongly support the use of available global data standards or the development of new standards in fields where none are available to ensure early alignment **- Ownership: Common**
- Where data cannot be standardised at inception, establish the regulatory requirements to confirm the validity of mapped data **– Ownership: EMA/HMA**
- Promote use of global open source file formats **– Ownership – common**

Almost without exception each of the subgroups raised the need for standardisation as a key pre-requisite in order to drive harmonisation across datasets, enhance interoperability, improve data quality and facilitate data analyses. However before summarising the recommendations in this area, it is helpful to define the term data standard as this is often used ambiguously. The overarching term data standard can be defined as *'a model to represent a data entity or series of entities and provides a mechanism to provide consistent meaning to data shared among different information systems'*. There are a number of terms which are commonly used which underpin the overarching term data standard which are described in Annex VI.

The need for standards was recognised many years ago and when required for regulatory purposes has driven global harmonisation. For instance, the data model and data elements of the Individual Case Study Report (ICSR) used for reporting of adverse drug reactions (ADRs) have been defined by the ISO ICSR data standard and the terminology within the ISCR for example the coding for ADRs is specified by MedDRA. Such clear specification of reporting requirements for ADRs means platforms such as EudraVigilance[10] contain extremely well structured information, although the completeness of the information within ICSR forms is dependent on the reporter and hence variable. Similarly while there are several data standards for clinical trial data[11], one of the most widely used standards for clinical trials were developed and are maintained by the Clinical Data Interchange Standards Consortium (CDISC) which provides standards to aid data collection at clinical investigation sites but also standards to structure and transmit data.

However many other datasets are not standardised partly due to the fact that most have evolved over many years and hence the data encompasses many technological developments. In addition, with the exception of the clinical trials data, data were not generated to support regulatory decision-making and hence the need to comply with strict quality guidelines. Thus data heterogeneity spans a continuum; consider genomics as an example where nearly 250 million genomes are currently available but while relatively well structured, much of the data is siloed by disease, institution and country, generated with different methodologies, analysed by non-standardised software, and often stored in incompatible file formats and consequently only a small percentage is linked. This situation is replicated over multiple

[10] http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000679.jsp
[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078665/

data sets in the big data landscape, particularly real world data and at the far end of the spectrum social media data. However, standardising the data is hard; much of data is unstructured and heterogeneous and this is especially true of social media data which is anticipated to account for much of the data volume increases in the coming years.

No single data standard will have the depth and breadth to be applicable to all data sets but the taskforce recommends that we should strive as much as possible to minimise the number of standards. Standards should be transparent, open to promote widespread uptake, globally applicable, maintained with an ongoing process for testing and revision and sustainable. It is therefore important to strongly support the use and maintenance of available data standards, developed for example by International Council for Harmonisation (ICH), Health Level Seven International (HL7), International Consortium for Health Outcomes Measures (ICHOM), International Organization for Standardization (ISO) and Clinical Data Interchange Standards (CDISC) and the development of standards where none are available e.g. complex and dynamic data such as proteomics, novel data sources such as m-health and immature fields such as epigenetics to ensure early alignment. A recent example of regulatory uptake and support of standards is provided by ISO IDMP[12], which aims to facilitate international identification of medicinal products. While the benefits are clear, implementation of standards is expensive and requires concerted and harmonised action and thus the challenges associated with standardisation should not be underestimated. There is a need for a common understanding of the overall vision and scope, a clear definition of the ultimate value and a well-formed plan for implementation. Moreover, sustainability of standards is challenging but will be enabled by widespread adoption. From a regulatory perspective, a prioritisation of efforts will be needed with an early focus on data most likely to impact on decision making in the near term.

On a global level it is important to ensure that extremely expensive and time consuming initiatives[1, 13, 14] do not pull in opposite directions but work together to achieve sustainable and global solutions. From a regulatory perspective, global co-operation is important as for many rare diseases and cancers or indeed rare adverse drug reactions there may only be a handful of cases worldwide and these data needs to be interoperable to derive meaningful insights. We need to be aware also that data mapping is expensive, may create assumptions around equivalence and there is always a fear of information lost during data transformation, so therefore standardisation of data at inception should be the goal. Where this is not possible, a clear framework to confirm the validity of mapped data for regulatory decision-making needs to be established e.g. following the implementation of common data models.

[12] https://www.ema.europa.eu/human-regulatory/research-development/data-medicines-iso-idmp-standards/substance-product-organisation-referential-spor-master-data
https://www.efpia.eu/media/25717/principles-for-the-implementation-of-iso-idmp-standards-for-eudravigilance-and-development-of-a-road-map-2014.pdf

[13] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[14] http://yosemiteproject.org/

6.3.2. Data quality

Core recommendation

Characterisation of data quality across multiple data sources is essential to understand the reliability of the derived evidence

(Supported by subgroup recommendations # 12, 13, 14, 29, 33, 35, 37b, 41, 42)

- Characterise and document data quality in a sustainable EU inventory.
- Establish minimum sets of data quality standards. Where possible, quality attributes e.g. compliance to GCP requirements should be integrated to facilitate selection of appropriate data sets for analysis.
- Implement data quality control measures.
- Establish a clear framework for the validation of innovative bioanalytical methods e.g. 'omics.

Ownership of the Action: EMA / HMA

To a large extent, data quality determines the validity of the evidence that can be reliably derived from a given dataset. As 'big data' are associated with multiple and often diverse quality issues wherever possible, data should be characterised and minimal quality standards should be defined for both the source data, any transformed/mapped data and the subsequent analysis. Information should also be provided about the range of applicability, quality control measures and limitations of (selected) data and these findings should be transparently recorded in a sustainable, accessible inventory. Whether such standards are ultimately acceptable in a regulatory setting will depend upon the context of use.

The creation of a clear framework for acceptable data quality is complicated by the fact that the level of required data quality attributes will be variable for different regulatory applications. Thus, a classification system defining minimal requirements depending on the intended regulatory purpose is required. The current EMA Patient Disease Registry Initiative[15] provides an example of how incorporating the needs of relevant stakeholders informs the development of minimal quality standards and data elements in order to facilitate downstream data harmonisation.

[15] https://www.ema.europa.eu/human-regulatory/post-authorisation/patient-registries

6.3.3. Data sharing and access

Core recommendations

The development of timely, efficient and sustainable frameworks for data sharing and access is required

Further support mechanisms are needed to promote a data sharing culture

(Supported by subgroup recommendations # 3, 5, 9, 13, 18. 33, 18, 33, 37b, 40)

- Strongly recommend the establishment of distributed data networks to facilitate data sharing of sensitive healthcare data **– Ownership: EMA/HMA**
- Develop guidance for robust data governance and data anonymisation to deliver systems which secures patient trust **- Ownership - Common**
- Establish disease-specific minimum data elements to enable harmonisation of data across for e.g. national disease registries **– Ownership: EMA/HMA**
- Promote mandatory sharing of the analysis arising from data sharing activities e.g. by publication or open sharing via data access platforms **– Ownership: Common**
- Promote the sharing of qualified models**– Ownership: EMA/HMA**
- Support the development of policy initiatives to drive a data sharing culture which is mutually beneficial for all stakeholders. **- Ownership – Common**
- Proactively drive and/or support data sharing platforms and initiatives**– Ownership: EMA/HMA**
- Require the submission of data management plans at the start of all data generation exercises **– Ownership: EMA/HMA**
- Establish accountability for users**– Ownership: Common**
- Development of common principles for data anonymisation to facilitate data sharing **– Ownership: Common**

Data sharing can be defined as the practice of making original health data available for secondary research purposes by other investigators; data may be shared in various formats and the process of data release can range from sharing under open access arrangements to sharing under controlled and restricted conditions with named individuals or healthcare sectors. However whenever feasible, data should be shared as openly as possible.[16]

Data sharing is motivated by the belief that sharing and integrating data across multiple datasets maximises its possible benefit by enabling potential insights to be derived which may not have been possible from a single dataset. In addition it prevents duplication of effort and also helps ensure patients are not subjected to procedures from which they will derive no benefit or to duplicative and unnecessary trials. As a result research funders, journal editors, governments and regulators are

[16] The scope of work included in this report did not include a consideration of the legal requirements to protect patient privacy when sharing data. As such the recommendations outlined in this report and in Annex III will need to consider at all times data protection and meet the requirements of the of EU data protection legislation, in particular Regulation (EU) 2016/679 (the General Data Protection Regulation) or Regulation (EU) 2018/1725, as applicable. This will be incorporated into the next phase of the work.

increasingly demanding that data generators, be they academics, healthcare professionals or industry, commit to meaningful data sharing practices.

Despite the recognised benefits of data sharing, multiple barriers are preventing its natural progression some of which are common across datasets. Firstly it is becoming progressively more challenging to share increasingly complex data from multiple sources in sufficient depth and detail so as to retain its utility and meet data protection obligations on a global scale. Robust data anonymisation offers a route for sharing healthcare data at an individual patient data level but the challenge is to determine what level of risk of re-identification is acceptable in order to deliver the potential benefits of data sharing. Global guiding principles and standards for data anonymisation are urgently needed to resolve this dilemma and find an appropriate balance and consistency of approach to derive the benefits of data sharing. Such work will form part of Phase 2 of Policy 0070 which tasks the EMA to review the most appropriate way to make individual patient level data available while complying with privacy and data protection laws; this work has started through the establishment of the Technical Anonymisation Group[17] but should be progressed as a priority.

It is recognised that data sharing requires informed and detailed prospective planning to deliver success. As such data management plans, which describe the life cycle for the data to be collected, processed and generated for a project, including the use of standards, and how ultimately it may be shared and made open should become a mandatory part of any study. This ensures that the budgetary planning for resources required to make data accessible is considered at the inception of projects.

To derive maximum benefit the taskforce emphasised that data needs to be shared at a sufficient level of detail. Data sharing platforms should mandate sharing of meta-data and as a pre-requisite for accessing data investigators should commit to upload the analysis derived from data shared via the platform. Agreement of minimal data elements for specific disease areas would additionally support harmonisation and pooling of datasets. It is notable that Europe has failed to define a clear path to enable sustainability of many previous data sharing efforts, particularly for observational healthcare data, and defining this should be a priority in the future. It must be appreciated that a data platform requires resources beyond the initial investment and must encompass ongoing funding to enable the continual update and validation of these dynamic datasets. A more centralised mechanism for funding infrastructure platforms across Europe may allow the provision of continued funding for those platforms which can demonstrate the greatest impact.

Data sharing is additionally hindered by a reluctance to share data in order to promote individual career ambitions or protect potentially commercially valuable information. Mandating data sharing activities will help in some sectors as demonstrated by Policy 0070, funders initiatives such as the Horizon 2020 Open Research Data pilot[18] and measures from journals to share data underlying published papers[19]. However additional policy initiatives are needed to truly promote a data sharing culture which is mutually beneficial for all stakeholders. Hence appropriate metrics for data sharing activities, accepted by funding bodies and academic institutions, need to be developed to assign recognition e.g. recognition for the timeliness and quality of data sharing, for the number of downloads or citations, follow on publications in addition to the development of additional impact metrics such as EMA qualification procedures. Undoubtedly meaningful academic recognition will encourage and facilitate data sharing. In addition given that many scientific journals already require the publication of genomic sequences behind scientific results it is the view of the task force that genomic sequences submitted as part of a regulatory application could be published in a similar fashion. Moreover these

[17] https://www.ema.europa.eu/human-regulatory/marketing-authorisation/clinical-data-publication/technical-anonymisation-group
[18] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[19] http://www.icmje.org/journals-following-the-icmje-recommendations/ ; https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html

should be shared (with appropriate data protection measures) so as to enable linkage to the disease and clinical outcome data with which they are associated.

6.3.4. Data linkage and integration

Core recommendation

Promote mechanisms to enable data linkage to deliver novel insights

Facilitate harmonisation of similar datasets

(supported by subgroup recommendations #6, 12, 15, 21, 34)

- Encourage sharing of raw data, associated meta-data and processed data to enable meaningful data linkage.
- Proactively engage with initiatives to map terminologies to facilitate data linkage and timely data access but ensure frameworks for consistent validation are simultaneously implemented.
- Support mechanisms to maintain up to date mappings across terminologies.
- Promote the inclusion of clinical outcome data relevant to regulatory questions in public databases.

Ownership of the Actions: EMA and HMA

The vital importance of data linkage between related datasets to provide additional insight not possible from single isolated datasets was a constant theme across all subgroups of the taskforce. This applies not only for databases within a subgroup e.g. how to integrate different registries or electronic health records described within the observational subgroup but also across subgroups e.g. linking clinical data with genomic/pharmacogenomics data and proteomic data and care settings e.g. primary, secondary and tertiary care. The standardisation approaches described earlier would significantly promote interoperability and drive harmonisation of data of the same type but it is also critical to develop linkages among disparate data sets. For example 'omics' data may predict a patient's response to a therapeutic intervention and thus minimise exposure of patients to inefficient or intolerable therapies, paving the way for personalised medicine in the future; however such insights can only be derived if 'omics data is linked to phenotypic outcomes. Currently clinical outcome data relevant to regulatory decision making e.g. data on efficacy or safety of treatments is only found sporadically in public databases limiting their value in a regulatory context; raising awareness of the need for data linkage of treatments and outcomes could be particularly beneficial.

Different questions will require linkage of data at different levels and require different data protection solutions. For most regulatory needs linkage at an individual patient level would ideally be required e.g. understanding the clinical outcome of an adverse drug reaction or to enable longitudinal follow up of a treatment. However there are scenarios where linkage of data at a population level may be sufficient e.g. standard of care at different disease stages across Europe or outcomes from vaccination programmes. To enable meaningful data linkage, sharing needs to move beyond simply sharing the raw data, to encompass associated meta-data which describes key characteristics about the data, e.g. sample type, disease stage, treatment, genomic mutation which will do much to promote meaningful data linkage.

Distributed datasets where personal identifiable data is retained within secure local storage but structured in such a way as to allow rapid interrogation seems the most likely solution to allow linkage of many datasets. However the development of such distributed databases needs to be supported by the simultaneous development of analytical approaches e.g. artificial intelligence (AI) approaches such as machine learning, able to derive insights across distributed data networks, scientific problems and tasks.[20].

Increasing linkage of healthcare data, especially if at an individual patient level, increases the risk of re-identification, may require agreement from multiple data owners and raises important ethical-legal issues. The consequences of this can often be restricted access for external stakeholders as is the case with many of the well linked Nordic registries. Investment in novel technological approaches for the management of patient level data which do not require the physical transfer of data e.g. DataSHIELD[21], block chain and homomorphic encryption and meets national and international data protection legalisation are urgently required.

6.3.5. Data analytics

Big Data analytics is a growing field of data science which combines methods from various disciplines including biostatistics, mathematical modelling and simulation, bio-informatics and computer science including data collection , data management, data-integration, data standardisation, machine learning and high-performance and specialised IT architectures and tools to extract knowledge and insights from data in various forms both structured and unstructured. This area was quickly identified as a key area to be addressed once the HMA/EMA Joint Big Data Task force was formed. As a response to needs arising in the other subgroups, the cross cutting analytics subgroup was formed early 2018, as common themes across subgroups needed a consolidated approach.

The below preliminary recommendations and reinforcing actions are based on the reports of the subgroups. Even at an early stage it is clear that in this fast moving field, continuous support will be required to access appropriate expertise to allow a critical appraisal of new methodologies as they arise. Formation of a standing advisory group is therefore recommended to explore the applicability of big data analytic methodologies, standards and IT architecture to support the development, scientific evaluation, supervision and monitoring of medicinal products. Secondly, validation of novel analytical approaches and the clinical relevance of the derived endpoints will be a key part in defining their acceptability especially for algorithms, which are continuously updated over time and deliver complex composite endpoints. In this area, the interpretability of the model results, defined as understanding how results are produced and having confidence that the model is performing accurately with respect to the desired objectives and scenarios, will be a key component. As such data platforms which incorporate bioinformatics applications addressing metadata documentation, standardisation, annotation and data management as well as providing open user-friendly algorithms and tools, and direct coupling to dedicated and performant statistical analysis increases the transparency of the analysis and are to be encouraged[22]. Utilisation of EMA Qualification Advice process will enable regulators to influence more mature approaches. Lastly, unstructured clinical information will continue to appear in textual clinical notes for many years to come. Thus, a document architecture standard is needed to enable the interchange of clinical notes and to facilitate the extraction of information using natural language processing techniques.

[20] See Section 4.3.5:Data analytic recommendations

[21] (https://www.ncbi.nlm.nih.gov/pubmed/25261970)

[22] See Section 4.3.3.:Data sharing and access

Core provisional recommendation*

Develop clear frameworks to enable the validation of analytical approaches to determine if they are appropriate to support regulatory decision making

Promote new analytical approaches for modelling of big data sets for regulatory purpose

(Supported by subgroups' recommendations # 4, 6, 11, 19, 20, 23, 42 and pending recommendations of the data analytics subgroup).

- Move the analysis to the data: actively support the development of novel analytical approaches (e.g. AI, machine learning) applicable across distributed data networks which do not require the physical transfer of data.
- Form an advisory group to:
  - explore the applicability of novel analytics methodologies to support the development, scientific evaluation and monitoring of medicinal products;
  - Explore the most suitable data standards and IT architecture and tools capable to enable the analyses.
- Promote the increased utilisation of scientific advice and the EMA Qualification Advice process to enable regulators to influence more mature approaches.
- Support, define and validate the definition of innovative outcome measures and other approaches which leverage additional dimensions from high-frequency or high-dimensional data.
- Make publicly available data analysis plans for all studies submitted for regulatory approval.
- Strongly support the exploration of novel analytics approaches such as natural language processing techniques to interrogate unstructured data.

Ownership of the action: EMA with HMA support

This is a particular challenge for European data which originates in multiple different languages and is influenced by local healthcare practices.

*Recommendations and reinforcing actions are provisional pending the full report and recommendations of the data analytics subgroup.

Provisional recommendations will be updated following the recommendations of the data analytics subgroup in 2019.

6.3.6. Regulatory acceptability of Big Data analyses

Core recommendation

Regulatory guidance is required on the acceptability of evidence derived from big data sources.

(Supported by subgroup recommendations # 10, 13, 14, 28, 29, 47*).*

- Identify the best format to enhance the agility of guidance development and revision in this fast moving field.
- Track concrete examples of procedures relevant to big data across the regulatory network to inform thinking.
- Establish pilot programmes to develop informal discussion on acceptability.
- Initiate pilot studies to better understand the evidence generated on efficacy/effectiveness and safety from emerging datasets.
- Mandate transparency and format around study reporting for regulatory submission to document datasets, protocol, tools and version used to promote reproducibility.
- Emphasise the need for outcome measures from novel data sources e.g. m-health devices to be reflective of a defined clinical benefit.

Ownership of the action: EMA with HMA support

The regulatory environment is changing. We are seeing an increasing number of innovative products that face challenges aligning with the traditional drug development pathway which creates additional uncertainties at authorisation which must be carefully managed post authorisation. In addition we undoubtedly will need to assess data from multiple new emerging data sources and as a regulatory network we must prepare for and understand this change in data generation and knowledge management. It is important that the need to maintain our evidentiary standards does not result in a reversion to the status quo and a failure to exploit the potential opportunities.

Today the process of generating evidence from big data sources is far from a straightforward, pre-defined journey from source data to actionable evidence. Uncertainties about the quality of the data, the models and the level of quality management used undermine the confidence in the validity and reliability of the evidence generated**.** Understanding how to reduce or understand the variability in the evidence generation pathway to increase trust in its ultimate product will increase regulatory acceptability and promote its uptake and utilisation. The actions outlined to date in the report, particularly increased standardisation and measures to understand and document data quality will be key steps along the road to regulatory acceptability.

Guidance is clearly needed but in fast moving fields it is necessary to identify the best format in order to enhance the agility of development and revision. Guidance should clearly state what should be reported and how and should be relevant to what is being presented through regulatory submissions. For example guidance may define the minimum quality requirements which should be addressed to cover data consistency, accuracy, reproducibility, representativeness and missingness along with the quality control and assurance measures in place to guarantee the data elements. For digitally captured data, quality measures would need to incorporate algorithms and the device parameters including sensitivity, specificity, accuracy and precision of the delivered measurement. As such through EMA Qualification Advice, opinions have already been provided on novel endpoints provided by wearables for utilisation in clinical trials[23], ingestible sensor systems for adherence[24], on novel biomarkers[25] and on data sources appropriate for regulatory decision-making[26]. Use of tools for tracking innovation in EMA procedures and products and business intelligence tools would inform the need for guidance in a particular area. The ultimate vision is to create a clear framework under which regulators could determine the potential acceptability of the evidence presented to them and to deliver a consistency and clarity of approach for external stakeholders to work within.

6.3.7. Medical devices regulation (MDR) / In vitro Diagnostics Regulation (IVDR)

Core recommendation

Ensure effective implementation of the new regulations for devices and in-vitro diagnostics (IVDs) associated with the use of medicinal products and monitor its impact in delivering safe and effective devices and IVDs

(Supported by subgroup recommendations # 28, 30, 31, 38*)*

- As a minimum, for innovative devices and those incorporating complex algorithms ensure effective coordination processes are implemented across multiple national notified bodies and national regulatory drug agencies to establish common specifications on analytical and performance requirements for similar devices and IVDs.
- Develop strong and systematic partnerships between notified bodies and regulatory agencies (medicine and medical device).
- Closely monitor the impact of the updated EU Medical Regulation to determine whether it meets the evolving needs.
- Harmonisation of European reporting of adverse events/incidents associated with companion diagnostic-IVDs (CDx).
- Implement mechanisms (e.g. reference labs) to quantify comparability of different tests (CDx) for the same biomarker.

Ownership of the action: HMA with EMA support

Medical devices and In-Vitro Diagnostics (IVD) / IVD-Companion Diagnostics (CDx) play an increasing role in the health care systems of Europe, both in isolation and increasingly in combinations with medicinal products or other advanced therapies. The pace at which the device / IVD-CDx sector is evolving is rapid, generating specific challenges for both notified bodies and regulatory agencies in maintaining the skills necessary to regulate the industry. While authorisation of the devices per se falls outside the remit of the taskforce, the development of the device may well have relied upon large dynamic datasets, the reliability of the outputs of the device may rely on algorithms that change constantly depending on the data input or the data that is ultimately generated through the device

[23] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/draft-qualification-opinion-stride-velocity-95th-centile-secondary-endpoint-duchenne-muscular_en.pdf
[24] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-ingestible-sensor-system-medication-adherence-biomarker-measuring-patient_en.pdf
[25] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-plasma-fibrinogen-prognostic-biomarker-drug-development-tool-all-cause_en.pdf
[26] https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-european-cystic-fibrosis-society-patient-registry-ecfspr-cf-pharmaco_en.pdf; https://www.ema.europa.eu/documents/regulatory-procedural-guideline/draft-qualification-opinion-cellular-therapy-module-european-society-blood-marrow-transplantation_en.pdf

may well be classified as big data. Increasingly there is a convergence between drugs and devices/IVDs which will undoubtedly influence the benefit risk of the medicinal products with which they are associated. As such the taskforce felt it appropriate to consider some of the key challenges associated with them within its work.

6.3.7.1. Medical devices

Brand new technological advances such as Apps for smart phones and tablets, wearable technology, AI-based algorithms provide many exciting new opportunities for revolutionising the health care sector both in diagnostics and treatment. There is currently an uneven capability across Europe for the assessment and regulation of these technologies. No knowledge sharing takes place on a routine basis, but may indeed be performed on an ad-hoc basis. The current European system with the revised legislation of 2017 has now introduced provisions to ensure that information on serious incidents or field safety correction actions for certain devices e.g. companion diagnostics are shared between authorities. However, it is not transparent at present how this will be achieved. At present adverse events are only reported de-centrally to NCAs when the issue is related to a device only. Therefore, the benefit harvested by sharing of information regarding potential adverse effects of medicinal therapies already in place in Europe is largely absent in the device area.

6.3.7.2. In-Vitro Diagnostics (IVD) / IVD-Companion Diagnostics (CDx)

Companion diagnostics are clearly moving into focus as part of the development of increasingly personalised treatments for more precisely defined patient groups. This has the benefit of limiting patient exposure to those medicines where the chance of success is improved (i.e. the benefit / risk relationship is improved). However, the prerequisite for this paradigm is that the diagnostic component performs as well as the therapy it accompanies. Fully validated tests with adequate sensitivity and specificity standards are therefore required; otherwise a false negative scenario may well evolve, where a patient is denied treatment with an otherwise effective drug simply because of an inaccurate companion diagnostic. Particular attention needs to be paid by the regulating authority to prevent such scenarios from arising.

A specific challenge has arisen regarding AI based devices and algorithms (machine based learning). By their very nature these algorithms are in constant change /evolution and the decision on how and especially when to evaluate these becomes apparent. FDA has recently approved AI-based algorithms for diagnosis of diabetic retinopathy[27] and detection of wrist fractures.[28] The consequences of a misdiagnosis or delayed diagnosis are all too apparent, and patient safety clearly demands sufficient regulation.

The Competent Authorities for Medical Devices (CAMD) have developed a comprehensive roadmap[29] to aid in the implementation of the new EU Medical Devices Legislation[30]. It is critical we achieve harmonised technical and clinical validation/performance standards in order to deliver consistent assessments across notified bodies. Linked to this is the need to harmonise the approach to risk management and safety reporting. Effective delivery of the roadmap will require efficient collaboration between multiple bodies and working groups both from within the medical devices network and external to it, including the European Commission and EMA. The roadmap lays the foundation but does not include details on who takes the lead on the multiple different activities and how a level of communication will be enabled which ensures that all relevant stakeholders are able to contribute to

[27] https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm
[28] https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm608833.htm
[29] https://www.camd-europe.eu/wp-content/uploads/2018/05/NEWS_171107_MDR-IVDR_RoadMap_v1.3-1.pdf
[30] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745

the assessments in an appropriate and timely fashion especially for areas in which there is joint responsibilities.

6.3.7.3. Summary

It is the recommendation of the task force that there is a need for a coordinated oversight of the authorisation of devices /IVDs/CDx many of which will require a very specific skillset. As devices/IVDs/CDx and medicines become increasingly interdependent, strong and systematic partnerships are needed between notified bodies, national medical device agencies and the medicines regulatory network to ensure the assessment considers the necessary requirements of the medicine to which the device is linked.

6.3.8. Skills and knowledge across the regulatory network

Core recommendation

Regulators must be equipped with the skills required for these emerging areas

(Supported by subgroup recommendations # 2, 22, 27, 30, 36, 46*)*

- Identify skills gap.
- Develop prioritised recruitment strategies.
- Establish a network of bio-informatics/biostatistics/analytics/data science expertise and excellence within the European regulatory Agencies.
- Develop strong bi directional collaborative links with academia to communicate regulatory needs and maintain an up-to-date knowledge in the regulatory network.
- Strengthen regulatory science networks.
- Implement training programmes.

Ownership of the action: EMA/HMA

The use of Big Data approaches and their analysis calls for new regulatory strategies and guidance to achieve their full potential. Specialised expertise in the field of data science which combines various disciplines such as biostatistics, mathematical modelling and simulation, bio-informatics and computer science will be required to allow an informed and critical assessment of regulatory applications in the future. There are a number of relevant EU level working parties already established including the Biostatistics working party, the Modelling and Simulation working party and the Pharmacogenomics working party within which some of this expertise currently sits. However as many of data science disciplines complement each other mechanisms need to be found to efficiently link experts when required to not only deliver an appropriate assessment of innovative methods, approaches and products, but also to support and maintain skills in the regulatory network. Relevant knowledge gaps revealed by the assessment of products must be identified on a regular basis and addressed by targeted recruitment strategies to ensure the operational capacity of the regulatory system.

Efficient bidirectional communication and collaboration with academic experts and centres of excellence should be established to maintain regulatory knowledge of current developments and innovations and to inform academic researchers about the requirements for regulatory acceptability of new methods and treatments. Such links will also inform the development of relevant and current open training programmes across a range of levels of expertise to build capacity and expertise across the regulatory network.

There should be strong support for development and establishment of regulatory science in the fields of bio-informatics, data science and personalised medicine. A focus on regulatory science during professional training to communicate regulatory knowledge will improve quality and accelerate development of innovative treatments in the future for the benefit of patients.

6.3.9. External communication and engagement

Core recommendation

Proactive regulatory engagement with external stakeholders relevant to the Big Data Landscape is needed in order to influence strategy and ensure regulatory needs are highlighted.

(Supported by subgroup recommendations #19, 24, 27, 31, 37, 39, 44*)*

- Close co-ordination of all big data related activities to reduce duplication of effort and enhance sustainability.
- Engagement with data generators/academics to highlight regulatory needs for data generation, recording of meta data, data standards and data analysis.
- Work to align thinking across all stakeholders to develop unified strategies.
- Support patient communication channels to increase awareness of the value of data sharing.

Ownership of the action: EMA/HMA

It is clear from recent landscaping work[31] that there has been significant investment in many aspects of big data and that this investment is set to continue with recent announcements such as the European Science Cloud, digital single market strategy, FAIR data and the Innovative Medicine Initiative's BD4BO programme. However, the restricted sustainability of many of the projects funded to date, in part due to the short-term nature of the funding and lack of centralised long term infrastructure funding has limited the downstream utility and impact. One notable exception to this which is relevant to Big Data, is the European Bioinformatics Institute, one of six sites of the European Molecular Biology Laboratory (EMBL) which serves as a good case study as to what centralised long term funding can deliver.

Engagement with multiple stakeholders will become increasingly important as sources of data generation change. For example by 2020 it is estimated that more than 80% of the whole genomes and exomes that are sequenced will be funded by healthcare systems as opposed to 20% as of today. This may facilitate linkage of genomic and healthcare data but conversely may complicate access. Similarly, the bulk of m-health data, which promises the ability to capture a more holistic picture of the patient, is generated by individuals not by institutions but may be owned by commercial companies such as Google and Apple. Thus, it is critical that communication channels with patients and healthcare providers raise awareness of the value of data sharing and develop network of trusts between multiple actors. Regulators need to be part of this of this conversation to ensure regulatory needs are recognised.

[31] https://bmjopen.bmj.com/content/8/6/e021864.long

Big data has a reproducibility challenge not only because the data sets are dynamic with unknown provenance but meta data is not always fully described which makes it very challenging to document the data and analytical journey. Agreement from stakeholders to describe their data in a comprehensive and standardised manner will significantly increase replicability but requires constant engagement with all relevant actors.

Standardisation at inception is far preferable to data transformation and mapping but if it is to meet regulator needs it is critical to engage with initiatives to ensure that the data generated at source meets needs as closely as possible.

7. Delivering a Big Data Roadmap

With the present report the Big Data Task Force concludes its work according to the mandate given by the joint HMA/EMA management groups (analytics subgroup work excepted – delivery: Q12019). The overarching conclusion is clear: much may be gained from the rational use of Big Data in a regulatory context for approval and monitoring of efficacy/effectiveness and safety of medicines, medical devices and combinations thereof. Indeed many future activities necessary for regulatory progress will not be possible without the use of big data technologies. AI technologies offer particularly promising advances in these fields.

It is however clear that without a systematic, coordinated and integrated European approach many of these advantages may not be gained. Challenges of great complexity remain to be solved particularly regarding data access, transfer, interoperability and data quality as outlined by the respective subgroups. Moreover the timescale over which these recommendations must be implemented is long and will require continual iteration and reconsideration as new developments and methodologies emerge. However tasks must be tackled in a sensible order to enable the regulatory system in Europe to contribute and support the exploitation of these data sources in the assessment of medicinal products. A high level roadmap is proposed in Figure 1 which suggests a number of steps on the road from big data to regulatory acceptability. The steps are not necessarily sequential, many are interdependent and all will require active and iterative communication between all stakeholders.

Notably the current reports sets out ‘what‘ needs to be addressed but ‘the which‘, ‘the how‘ and ‘the when‘ will need further work and as such the mandate of the Task Force has been extended. Importantly as viewed through the regulatory lens, datasets are not equally heterogeneous in structure and quality and hence some will be more immediately relevant for regulatory decision making. A prioritisation and focusing of actions, that is ‘the which‘, will be required, which should build on ongoing actions, to deliver outputs in a timely manner and the extended mandate reflects this need. The scope of work is large and in some areas we are already moving in the right direction but not in a consistent and consolidated way and we therefore need to guard against reverting to the status quo. Rather, when challenged with new scientific and technological possibilities we should engage in order to ensure we have the capability and capacity to analyse, interpret and profit from the data generated. In this way we will improve our decision making and enhance our evidentiary standards.

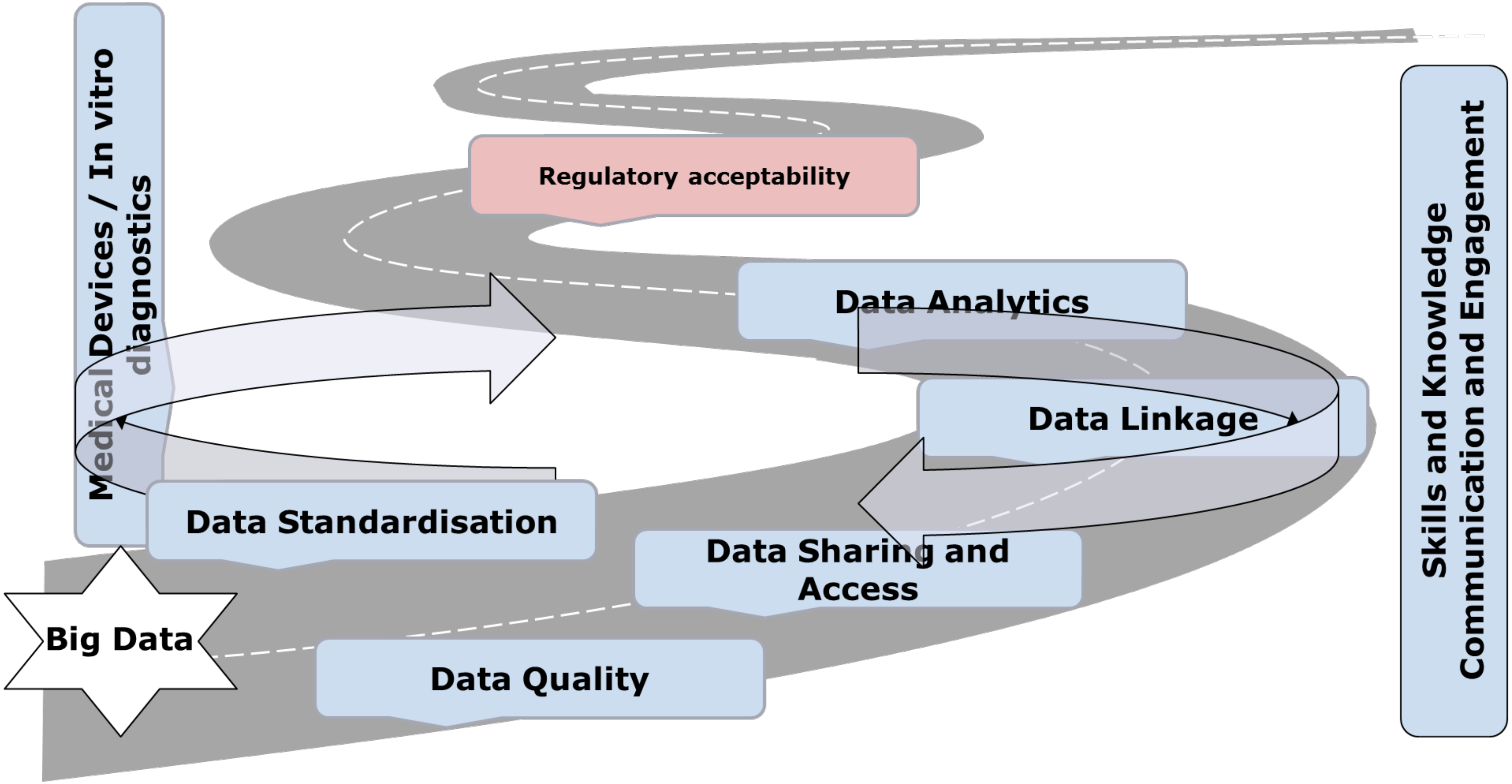


Figure 1: The Road to Regulatory acceptability: an integrated strategy reflecting core recommendations to support the use of Big Data in the assessment and monitoring of medicinal products in Europe. The individual steps are not necessarily sequential, may not be required across all datasets, many are interdependent and all will require active and iterative communication between all stakeholders.

8. Annexes

8.1. Annex I: Subgroup membership

Clinical Trial and Imaging Subgroup

- Vesa Kiviniemi, FI (subgroup lead, from October 2017);
- Per Fuglerud, NO (subgroup lead until June 2017);
- Aldana Rosso, DK (since October 2017);
- Zsuzsanna Cserjes Szabone, HU;
- Martin Nyeland, DK (until October 2017);
- Ada Georgescu, RO (until January 2018).

Observational data subgroup

- Aldana Rosso, DK (subgroup lead, from September 2017);
- Martin Erik Nyeland, DK (subgroup lead, until September 2017);
- Alexandra Pacurariu, EMA;
- Alison Cave, EMA;
- César Hernandez Garcia, ES;
- Katherine Donegan, UK;
- Marjon Pasmooij, NL.

Genomics subgroup

- Marjon Pasmooij, NL (subgroup lead);
- Didier Meulendijks, NL;
- Dieter Deforce, BE;
- Hans Ovelgönne, NL;
- Renate König, DE.

Spontaneous ADR subgroup

- César Hernandez García, ES (subgroup lead);
- Luis Pinheiro, EMA;
- Miguel Ángel Maciá, ES;
- Roxana Stroe, RO (until January 2018);
- Ada Georgescu, RO (until January 2018);

- Roxana Dondera, RO (until January 2018);
- Zsuzsanna Szabóné Cserjés, HU (until June 2018).

Data analytics subgroup

- Paolo Alcini, EMA (subgroup lead);
- Gianmario Candore, EMA (subgroup co-lead);
- Hans Ovelgonne, NL;
- Luis Pinheiro, EMA;
- Mateja Sajovic, SI;
- Kevin Horan, IE (until November 2017);
- Panagiotis Telonis, EMA;
- Antti H Hyvärinen, FI;
- Marek Lehmann, EMA.

Bioanalytical Omics subgroup

- Renate König, DE (subgroup lead);
- Alison Cave, EMA;
- Didier Meulendijks, NL;
- Mark Goldammer, DE.

Social media/M-health data subgroup

- Katherine Donegan, UK (subgroup lead);
- Hans Ovelgonne, NL;
- Gavril Flores, MT;
- Per Fuglerud, NO (until October 2017);
- Ada Georgescu, RO (until January 2018).

Survey design and analysis

- Kelly Plueschke (EMA).

Secretariat functions are provided by Randi Munk-Jakobsen (DK). Tina Engraff (DK) provided Secretariat functions until October 2017.

8.2. Annex II: List of Regulatory Decisions

8.2.1. Pre-authorisation phase

- Orphan designation;
- Certification and inspection of laboratories – GLP compliance (pre-clinical studies);
- Authorisation of clinical trials – regulatory and ethical;
- Amendments of clinical trials;
- Corrective measures;
- Paediatric investigation plan (PIP);
- PRIME (priority medicines) eligibility and early dialogue between the companies and regulators;
- Innovation support by national innovation offices (EU Innovation network);
- National and EMA scientific advice, protocol assistance;
- Request for eligibility to the centralised procedure;
- Appointment of the Rapporteur(s);
- Pre-authorisation inspections and quality assurance (GMP, GLP, GCP);
- Quality control: sampling and testing of medicinal products;
- ATMP certification;
- Classification of medicinal products;
- Request for accelerated assessment;
- Assessment of the national manufacturing licenses (hospital exemption for ATMPs);
- Compassionate use opinions;
- Biowaivers for bioequivalence trials.

8.2.2. At the time of marketing authorisation

- Evaluation of marketing authorization (efficacy, safety, quality, RMP);
- Request for maintenance of the orphan designation;
- Evaluation of market and data exclusivity;
- Decision on orphan similarity;
- Approval of national translations of the product information;
- Decision on the conditions of the marketing authorization: standard, conditional approval, authorisation under exceptional circumstances;
- Decisions on risk management plan, PASS and PAES.

8.2.3. Post-authorisation phase

- Evaluation of variations and notifications (N, type IA, type IB, type II);
- Handling of applications for transfer of marketing authorisation;
- Evaluation of applications for extensions of marketing authorization;
- Evaluation of annual renewals (conditional MA);
- Annual re-assessment (MA under exceptional circumstances) 5 year renewal;
- Evaluation of +1 year additional market protection for new therapeutic indication;
- Assessment of post-authorisation safety studies (PASS);
- Assessment of post-authorisation efficacy studies (PAES);
- Referral procedures;
- RMP assessment;
- Safety monitoring, e.g. signal detection and assessment of PSURs;
- Authorisation of wholesale distributors;
- Post-authorisation inspections and quality assurance (GMP, GLP, GDP, GCP);
- Quality control: sampling and testing of medicinal products;
- Assessment of shortage situations (criticality assessment etc.);
- Assessment of the implications of cessation of marketing authorization;
- Monitoring of advertising of medicinal products.

8.2.4. Other post-authorisation decisions (mostly national and often performed by other body than a regulator)

- Health technology assessment (HTA);
- Pricing and reimbursement decisions.

8.3. Annex III: Table of Recommendations from the Subgroups

Subgroup Recommendations

Colours reflect initial prioritisation: green – top priority: recommendations reflect action which reflect urgent actions or which are required for other actions to proceed; blue - medium priority; represent actions which either are dependent on other prior actions or which reflect a lower priority ; grey – low priority: often actions which related to immature fields or which validate other activities. A further prioritisation and focusing of actions will be performed by the taskforce during the next phase of work.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
Clinical Trial and Imaging Subgroup Recommendations				
1	Data standardisation activities are critical to increase data interoperability and facilitate data sharing.	**Agree on data formats and standards for regulatory submissions of raw patient data.**	• Strongly support the use of available global data standards and alignment with other regulatory bodies to facilitate global clinical data sharing e.g. CDISC and ISO IDMP. • Encourage use of open source data formats global standards. • Establish guidelines for use of other types of data types such as DICOM for images (see recommendation no. 5) relevant to regulatory submissions.	Agreement on formats and standards.
2	Establish direct access to individual patient level data during review of the marketing authorisation	**The European regulatory network should have direct access to IPD during assessment of a marketing authorisation.**	• Agree on data format for regulatory submissions of IPD (see previous recommendation). • Establish a mechanism for storage and access to IPD for regulatory submissions. • Increase capacity and skills for analysis of patient level data at NCAs.	A system for direct and timely access to patient level data in the context of regulatory submissions.
3	Sharing of clinical trial data submitted for regulatory assessment	**Support policy for systematic data sharing of clinical trials.**	• Support for Phase 2 of Policy 0070 which seeks to determine the most appropriate mechanism to share IPD.	Agreement on the mechanism of sharing IPD which complies with privacy and data protection laws.
4	Exploiting images to inform regulatory science.	**Imaging expertise and reading capacity should be established for regulatory**	• Support open file formats and data standards for imaging data. • Establish centralised co-	Increased utilisation of images to support regulatory decisions.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
		purposes.	ordination of expertise in reading at regulatory agencies. • Support pilot studies to define innovative outcomes from imaging data. • Support initiatives to determine the validity of computer aided evaluation of images.	
5	Demonstration of value of data sharing	**To promote a data sharing culture it is essential to demonstrate the value of clinical data sharing for medicines development.**	• Support pilot studies demonstrating the value of clinical data sharing with regulatory relevance. Possibilities include: identification of safety signals, product class comparisons, indirect comparisons of closely related medicinal products.	Demonstration of the value of clinical data sharing in medicines regulation.

Observational Data Subgroup Recommendations (Electronic Health Records)

6	Inconsistent availability of healthcare data from secondary care	**Mechanisms are required to drive the, standardisation and access to secondary care data.**	• Proactively support approaches to improve the linkage of primary and secondary healthcare data. • Proactively support pilots for areas where data is lacking e.g. linkage of paediatric data across specialist paediatric centres. • Encourage standardisation of terminologies across care settings to facilitate linkage. • Create an inventory of reliable sources of secondary care data hosted on the ENCePP website. • Explore natural language processing techniques to interrogate unstructured data.	Increased access to data from secondary care.
7	Representativeness of observational data of European population	**Development of data sources in European member states which do not currently provide access to electronic**	• Encourage the development of electronic data sources in member states currently underrepresented in	Increase in number of countries with electronic healthcare databases that can be accessed to

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
		health records for observational research.	multidatabase studies.	support regulatory decision making.
8	Timely access to pan European healthcare data	**Sustainable mechanisms for combining healthcare data across Europe should be implemented.**	• Strongly support the establishment of distributed networks of datasets to improve timely access to data. • Where networks utilise a Common Data Model: – ensure the impact of transformation of data on the evidence generated is understood; – define the regulatory use cases for which distributed data networks dependent on a CDM would be acceptable; – ensure the maintenance of up-to date mappings as new data elements are introduced. • Support the development of robust data governance mechanisms to ensure data privacy obligations. • Emphasise the need for a sustainable solutions.	The speed of real world evidence generation across multiple datasets.
9	Multiple coding systems to record exposure and outcomes from medicinal products	**Increase the consistency of recording information on exposure to medicines including indications for use, product, dose and route, duration. Increase the consistency of recording of outcomes.**	• Support the implementation of ISO IDMP standards within electronic health records. • Support the mandatory recording of indications. • Support the mandatory recording of outcome measures including cause of death.	Increase in the consistency in the recording of exposure to medicines and utility of RWD.
10	Acceptability of RWE for regulatory decision making	**Development of a framework to articulate for what questions and contexts RWE may be acceptable across the product life cycle.**	• Create in depth characterisations for each data source to document strengths and limitations across a broad range of use cases. • Develop robust validation measurements to understand and document the validity of EHRs for regulatory questions.	Increase in the use and value of RWD to support regulatory decisions across the product life cycle.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			• Encourage use of EMA qualification procedures to document the utility of specific data sources for regulatory decision making. • Promote transparency of reporting which should include a clear justification of database choice, study design and subsequent protocol changes. • Support recording of all protocols in the EU PASS register and the publication of all study findings irrespective of outcome.	
			• Initiate pilot studies to compare the evidence generated on efficacy/effectiveness through both RCTS and observational data sources for the same question and endpoint in a high enough number of cases to determine the statistical agreement in terms of support for decision making.	
11	Improve the integration of new datasources	**Mechanisms should be developed to integrate new data sources with EHRs.**	• Support the development of standard terminologies and methodologies to enable the incorporation of data from novel data sources e.g. m-health, PROM in a consistent manner.	Increase in the availability of consistent information of lifestyle factors and PROMs in EHRs.

Observational Data Subgroup Recommendations (Patient Registries)

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
12	Implementation of common core data elements in registries Specify data quality attributes for data standards	**Harmonisation of data elements, standards, terminologies and quality attributes to improve data interoperability.**	• Faciliate agreement by registry holders on common core data elements to be collected by all registries in a given disease area. • Contribute and support the definition and inclusion of data elements relevance for medicines evaluation e.g. ADRs, co-morbidities,	Publicly accessible list of the common data elements (with their definitions) collected by registries in a given disease area. Increased use of registries in regulatory submissions.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			• Where possible common data standards and coding systems should be used. • Establish minimum set of data quality attributes acceptable for regulatory purposes across multiple disease areas.	
13	Promotion of the use of registry data for regulatory decision making	**The sharing of information between registries within a disease area should be encouraged.** **Implement measures to increase the acceptability of registry data for regulatory decision making.**	• Develop guideline for data sharing mechanisms. • Develop policy for access to and transparency of registry data. • Encourage multicountry registry platforms to apply for an EMA qualification opinion. • Promote and sustain the incorporation of disease registries within the ENCePP inventory of data sources.	Technical guidelines to support harmonisation of registries across Europe. Number of registries with complete information in the ENCePP inventory of data sources. Number of registry-based studies used in regulatory assessments. Increased use of registry data in authorisation applications and in post authorisation safety and effectiveness studies.
			• Agreement on appropriate quality of life and patient reported outcome measures to allow inclusion in registries in a given disease area. • Support patient awareness measures on the need for systematic collection of information on disease, treatments and outcomes in particular disease areas, especially rare diseases.	
14	Methods for interpretation of data	**Provision of guidance on accepted methods in registry-based studies with different purposes, e.g. monitoring of product utilisation, safety, efficacy, effectiveness.**	• Support continued update and use of ENCEPP guidelines. • Monitor the use of registry studies in regulatory decision making to assess the impact of quality standards and methodological guidance.	Increased value of registry studies in regulatory submissions.

Observational Data Subgroup Recommendations (Drug Utilisation Databases)

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
15	Limited availability of hospital drug	**Initiatives are required to increase**	• Support mechanisms to link primary and	Increased access to and use in-hospital

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
	prescribing	**access to hospital prescribing.**	secondary care prescribing. • Support development of smaller networks of hospitals, for example of paediatric specialist hospital, where prescribing could be consolidated.	prescribing data
16	Multiple coding systems to record exposure to medicinal products	**Increase the consistency of recording information on exposure to medicines including product, dose and route.**	• Support the implementation of ISO IDMP standards. • Implement mandatory recording of indications for use.	Increased consistency in the recording of exposure to medicines. Reliable verifiable linkage with community dispensing records.
17	Access to drug utlisation datasources	**Increase knowledge of the availability of drug consumption data**	• Creation and maintenance of a European inventory of drug utilisation data sources.	Increased availability of drug utility data.
18	Conduct of multi country drug utilisation studies	**Increase speed and quality of multi-country drug utilisation studies to optimally support signal assessment within pharmacovigilance.**	• Support the development of guidance by the International Society of Pharmacoepidemiology. • Support continued update and use of ENCEPP guidelines in this field.	Increased number of multi-database studies in Europe Number of multi-database studies registered in the EU PAS Register.

Spontaneous ADR Subgroup Recommendations

19	Data analytics	**Evaluate new analytical tools, such as forecasting and machine learning, that leverage increased dimensions of data (spatial-temporal, other variables in case reports, meta-data).**	• Promote the development of a system of oversight and tracking of innovative methods in signal detection (and any other use of pharmacovigilance data). • Strengthen the current processes that determine research priorities, track them and harvest EU-wide regulatory science skills to explore new analytical tools (e.g. PRAC's SMART Methods). • Boost engagement with key researchers in academia and other stakeholders. For that	An increased capacity and efficiency to analyse ADRs utilising novel analytical techniques.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			purpose the following could be considered: – Periodically publishing results from tracking of innovative methods and research priorities for the EU-regulatory system; – Host symposia dedicated to showcasing novel methods and research initiatives from stakeholders. – Host a competition with EudraVigilance data to improve signal detection methods. This would involve publishing a set of data from EV (anonymised) and setting research objectives. – Test output of new analytical procedures.	
20	Data content	**Explore how to implement natural language processing to improve efficiency, data management, quality and free expert reviewer time.**	• Investigate how to harvest the potential of automation in the EU-regulatory system particularly to improve the collection of data and its quality. • Facilitate the structuring of unstructured data (e.g. extracting relevant data from narrative fields such as case narratives, medical notes). • Assess the benefits and risks of increased automation.	An increased efficiency and capability to mine ADRs.
21	Data linkage	**Invest in methods to link pharmacovigilance data sources with other real world clinical and non-clinical data sources.**	• Develop parameters to enhance data linkage. • Ensure the maintenance of up to date mappings between MedDRA and coding terminologies used in observational data. • Perform pilot studies to determine how EV data could be linked with for example chemical structural data to	Increase in the accuracy numbers of reports. Increase understanding of the biological mechanism of ADRs. Increase in the ability to predict ADRs.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			enhance mechanistic understanding e.g. information on target proteins and off-site ample structural targets may help gain new insights. - Develop metrics to understand when and for which data such linkages provide greatest additional value. - Where appropriate foster the use of open source analytical software as a commitment to open science and to facilitate accessibility of research and analysis to all stakeholders.	
22	Skills and knowledge across the network	**Develop the capacity of the European regulatory network to assess new analytical approaches.**	- Recruitment of required expertise to enhance expertise within the regulatory network.	Increase skills and expertise of the regulatory network.

Social Media and M-Health Data Subgroup Recommendations (Social media)

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
23	Pharmacovigilance and signal detection	**Build on existing research on the use of social media data for providing insight into the identification of adverse events.** **Evaluate how social media can be used to monitor the safety and effectiveness of medicines.**	- Focus on specific areas e.g. quality of life, exposure during pregnancy, abuse/misuse, to understand how social media may contribute useful data. - Support further research into new analytical methodologies, including machine learning approaches to streamline the identification of relevant data. - Investigate how a wider range of social media data sources particularly patient forums may contribute to pharmacovigilance activities. - Contribute to research on, if and how social	To deliver an enhanced state of the art international pharmacovigilance system facilitating the rapid and robust identification of safety concerns.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			media reports can be integrated with other vigilance data sources. • Actively promote a coordinated, transparent, and collaborative approach to future research in this field involving researchers and organisations with the right scientific and technological expertise.	
24	Communication	**To actively research the use of social media for the communication of regulatory information.**	• Understand how behavioural science can contribute to effective messaging of regulatory recommendations on the use of medicines via social media to ensure changes in clinical practice. • Explore how the use of social media may facilitate patient recruitment into clinical studies. • Measure impact of communications in a qualitative and/or quantitative way. • Share experiences across the network on the use of social media by regulators for communication. • Consider potential reputational risks and best practices for engaging in discussion on social media.	To increase effective safety messaging, clinical management, and self-management.
25	Data access and use	**To ensure ethical and privacy issues on access to social media data are carefully addressed.**	• Support the development of guidance on the ethical and legal implications of using social media data.	Availability of appropriate guidance.
26	Data access and use	**Identify opportunities for gathering data from social media platforms.**	• Identify opportunities for regulators to access data from social media companies or to work with specific platforms to gather or stimulate new qualitative and quantitative patient reported data and take forward collaborations where appropriate.	Increased access to relevant patient-centric data.
27	Skills and knowledge within network	**Equip regulators with the new skills required for this emerging area.**	• Ensure there is sufficient expertise and capacities within the regulatory network.	Increased value of social media data within pharmacovigilance.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			• Support collaboration with academic and private organisations on the development of innovative approaches on the use of social media in pharmacovigilance practices.	
Social Media and M-Health Data Subgroup Recommendations (m Health)				
28	Validation of data coming from m-Health devices (including clinical trials)	**Facilitate the use of m-Health devices to record the efficacy and safety of medicines.**	• Map the different types of m Health data against their potential uses to define the extent and type of validation required from a regulatory perspective. • Use this mapping to determine (i) when specific guidelines are required and (ii) to what extent validation could be co-ordinated, (iii) at what level of data granularity and (iv) how regulators could support independent testing. • Ensure endpoints from m health apps are reflective of a defined clinical benefit that is relevant and important to the daily life of a patient.	Proactively defining expectations should deliver an increase in the validity of data submitted in regulatory submissions.
29	Collaborative working	**Develop an advisory board on m –health.**	Use this group to: • Support learning of medicines regulators on technological capability, data quality, analytical methodologies etc. • Help understand where m-Health technologies could have the greatest impact. • Support case studies that can be used to inform practice: • Help develop best practice guidelines and establish where data are fit for purpose. • Understand how apps can contribute to effective messaging of regulatory	An increase in the safe and effective use of mHealth technologies.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			recommendations on the use of medicines to ensure changes in clinical practice. • Identify new challenges and areas for future focus including the use of apps to understand patient preferences.	
30	Medical devices regulation	**Support effective regulation of m-Health devices used to generate data for medicines regulators.**	• Develop strong and systematic ties between device regulators and regulatory agencies in order that the different regulatory frameworks can operate in a complementary way. • Support upskilling of expertise and capacity at notified bodies and regulatory network to ensure they can robustly critique rapidly evolving devices and algorithms.	A co-ordinated device regulatory system with a high level of competence to ensure data quality and reliability of devices.
31	Collecting pharmacovigilance data and implementing risk minimisation	**Support effective vigilance practices using state of the art m-Health technology**	• Continue to develop apps for directly gathering data from patients on adverse events and encourage their wider use in real world and study settings. • Investigate how apps and other m-Health devices might be used by patients to support risk minimisation and optimisation of their use of medicines and where regulatory guidance on the collection, validation, and analysis of data is required.	An increase in the value of m-Health within pharmacovigilance.
32	Pharmaco-epidemiology studies	**Promote the use of m-Health technology to support effective post-authorisation studies.**	• Support case studies of m-Health technologies in order to better understand how these technologies could increase the strength of post-authorisation studies. • Establish standards for consistent data collection across apps.	An increase in the use and value of m-Health data within post-authorisation research.

Genomics Subgroup Recommendations

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
33	Sharing of genomic data (under a	**Stimulate public sharing of genomics**	To facilitate the sharing of genomics data from pivotal	Increased sharing of genomic data

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
	general recommendation of promoting a data sharing culture)	**and clinical trial data.**	clinical trials the following actions are required: – Define conditions, for sharing of genomic data including data anonymisation, minimal data elements, sharing of raw data in addition to processed data, mechanisms of access and security, and informed consent. – Ensure sharing of meta-data relevant for regulatory questions i.e. descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. – Recommendations on ethical issues unique to genomics e.g. familial issues, secondary incidental findings.	
34	Standardisation and data linkage	**Optimise data sharing and linkage of phenotypic and/or treatment parameters to genomics datasets.**	• Promote the use of harmonised open data file formats to improve sharing of genomics data and/or clinical outcome data linked to genomics data. • Promote linkage of relevant parameters (e.g. adverse events, primary efficacy outcomes) to the genomics dataset upon marketing authorisation application. • Promote interoperability of genomics data platforms. • Support pilot studies linking genomics data to clinical outcome data from different studies (efficacy/safety).	Increased linkage of genomic data to the key clinical parameters.
35	Evidence generation requirements	**Establish requirements regarding data quality for regulatory submissions.**	• Establish a working group to determine requirements for data quality, data standards, analytical methodologies, etc. • Initiate global collaboration regarding setting the standards for data quality	Define data quality standards/ requirements to ensure reliability of the analyses performed on big data sources.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			requirements.	
36	Skills and knowledge within the network	**To address the knowledge/ expertise gap across the European regulatory network to ensure big data applications can be reliably assessed.**	- Document the gap in knowledge/ expertise within the network. - Initiate training and recruitment to increase capacity.	Increased regulatory capacity for the assessment of genomics data.
37a	Need for regulatory guidance	**In fast moving scientific areas there is a need for faster and more agile regulatory guidance.**	- Identify the best format by which guidance can be developed in this rapidly changing field.	Improve the agility of guidance generation.
37b	Need for regulatory guidance	**Provide guidance for industry/academia in the use of big data in regulatory processes where current guidance is limited.**	- Provide guidance on validation of advanced genomics methods (e.g. sequencing) and on standardisation of data processing and analysis techniques, data standards and (open) data file formats. - Publish regulatory recommendations on genomics data sharing for regulatory submissions.	Increased consistency and quality of genomic data reporting in regulatory submissions.
38	Medical devices regulation	**To ensure effective regulation of genomic diagnostic tests which are associated with the use of medicinal products.**	- Ensure strong co-ordination to ensure consistent decisions on similar devices across Europe. - Develop strong and systematic ties between medicines and device regulators in order that the different regulatory frameworks can operate in a complementary way. - Improve evidence base for (label compliant) genomic testing in clinical practice complemented by clinician education and decision support. - Harmonisation of European reporting of adverse events/incidents associated with companion diagnostics across relevant authorities. - Encourage method qualification in order to understand comparability of	A device regulatory system with a high level of competence to ensure data quality and reliability of devices which influence medicinal product prescribing or ongoing monitoring are fit for purpose.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			different tests for the same endpoint. • Ensure review of the implementation of the updated EU Medical Directive regulation to determine whether it meets the evolving needs.	
39	Availability of clinically meaningful genomic information	**To minimise inconsistency in the availability of clinically meaningful genomics information in the SmPC for medicines.**	• Establish mechanisms for timely update of genomics information in relevant SmPCs. • Explore other ways for publishing curated, clinically meaningful genomics data in close collaboration with relevant academic stakeholders	Increase the availability of clinically meaningful, curated, up to date, genomic information relevant to medicines.
40	Demonstration of value	**Demonstrate the value of genomics/clinical big data analyses for medicines regulation.**	• Support the analysis of systematically gathered genomics data coupled to pivotal clinical trial data (efficacy/safety) e.g. by performing a pilot study in oncology. • Promote pilot studies to demonstrate the added value of genomics data for pharmacovigilance purposes e.g. by investigating the feasibility and the additional value of requesting MAHs to retrieve and submit genomics data from patients who experience severe/fatal ADRs.	Demonstration of the utility of Big data in medicines regulation.
	Bioanalytical ’Omics Subgroup Recommendations			
41	Quality [samples and documentation]	**Guidance should be provided on acceptability on Big Data sets to support regulatory decision making.**	• Quality attributes of big data sets need to be defined by regulators including appropriate data (file) formats and data standards. • Quality attributes should be included in order to allow appropriate selection, analysis and interpretation of data sets. • Define meta-data necessary for regulatory needs.	Increase in the number of validated/ qualified ‘omics’ Big Data biomarkers.
42	Bioanalytical	**Clear guidance**	• Standards for method	Increase in the

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
	method validation	**should be provided for the validation of bioanalytical methods suitable for the complexity of 'omics' techniques.**	validation should be specified by / or in close collaboration with relevant competent authorities. • Quality relevant aspects of bioanalytical method validation as well as data processing, data analysis and interpretation should be addressed in specific recommendations.	number of validated/ qualified 'omics' Big Data biomarkers.
43	Comprehensive-ness of available data sets ['bioanalytical omics']	**Assess the completeness of available data and the potential impact of missing data / information.**	• Establish a suitable framework specifying the conduct of bioanalytical 'omics' big data analytical approaches.	Increase in the number of validated/ qualified 'omics' Big Data biomarkers.
44	Supporting the harmonisation of data (file) formats	**Harmonisation of the used data (file) formats.**	• In order to establish an Open Data Mandate it is crucial to identify or develop open source file formats which include the relevant data and information (e.g. relevant metadata). • Regulatory agencies should advise which data file formats and / or attributes of data formats are acceptable for regulatory purpose.	Increase in the number of available, relevant and harmonised 'omics' big data sets acceptable for regulatory decision making.
45	Strengthening the development and harmonisation of data standards	**It is encouraged to minimise the number of data standards used.**	• Suitable and appropriate data standards should be identified and if necessary, adapted for the use in big data approaches. • Data standards should be platform-independent, appropriately validated and freely available.	Increase in the number of available, relevant and harmonised 'omics' Big Data sets acceptable for regulatory decision making.
46	Knowledge /expertise gaps within the European regulatory network	**To ensure appropriate assessment of regulatory submissions expertise in various disciplines (e.g. mathematical modelling and simulation, bio-informatics and computer sciences) will be needed.**	• Recruitment of appropriate expertise where none exists in the regulatory network. • The required capacities should be trained through a focused training programme on a European level. • Development of case studies to train and strengthen the capacities of the European regulatory network in the fields of computer science including data-	Increase in the number of competent assessors.

#	Topic	Core Recommendation	Reinforcing Actions	Evaluation Criteria
			integration/machine learning and high-performance computing.	
47	Regulatory recognition of clinical relevance and prognostic value of omics	**Regulatory agencies should clearly articulate what evidence is acceptable for proteomics in order to support regulatory decision making, highlighting their value as prognostic markers.**	• In line with guidelines developed for genomics, similar guidelines need to be developed for bioanalytical 'omics. • Regulatory guidance/advice should be provided via Qualification of novel methodologies process.	A framework of relevant guidance documents for regulatory use of big data ('omics') approaches. This should be accompanied by targeted and qualified scientific advice for particular projects and scientific questions.

8.4. Annex IV Survey questionnaires for NCAs and industry

https://www.ema.europa.eu/documents/other/hma/ema-joint-task-force-big-data-survey-national-competent-authorities_en.pdf

https://www.ema.europa.eu/documents/other/hma/ema-joint-task-force-big-data-survey-pharmaceutical-industry_en.pdf

8.5. Annex V: Surveys results

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

8.6. Annex VI – Definition of data standard terms

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

Standard terminology – is a set of “terms” that are shared, unambiguously understood and used among users to represent specific data elements in a database. Examples include SNOMED (Systematic Nomenclature of Medicine)[32], IDC-9 and IDC-10[33], MedDRA (Medical Dictionary for Regulatory Activities).[34]

Measurement Terminology – provides a standardisation of units to express “quantities“ in the same manner and when not possible (due to different jurisdiction) to have clear and unambiguous unit conversation rules. Universal principles for the expression of measurements have been defined by ISO 31, ISO 1000 and ISO 80000 series of standards, which implement the International System of Units (SI) defined by the General Conference on Weights and Measures.

Data model –an abstract representation which organises data fields in a relational manner to define the relationships between them and to identify how they relate to the characteristics of the real “objects”. When this representation becomes widely applied, shared and accepted by stakeholders, it may become a standard data model e.g. ISO/CEN. A data model is made of fields which can be filled using free text, standard and/or measurement terminologies.

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

An **electronic messaging standard** defines an electronic format to exchange a set of data fields in an un-ambiguous and interoperable way between stakeholders. In simple words this represents a way to encode data elements (including sequencing and error handling) to enable the transmission of data from one database to another.

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

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