# Big Data – How to Realize the Promise

Alison Cave[1,*], Nikolai C. Brun[2], Fergus Sweeney[1], Guido Rasi[1] and Thomas Senderovitz[2] on behalf of the HMA-EMA Joint Big Data Taskforce

**The increasing volume and complexity of data now being captured across multiple settings and devices offers the opportunity to deliver a better characterization of diseases, treatments, and the performance of medicinal products in individual healthcare systems. Such data sources, commonly labeled as big data, are generally large, accumulating rapidly, and incorporate multiple data types and forms. Determining the acceptability of these data to support regulatory decisions demands an understanding of data provenance and quality in addition to confirming the validity of new approaches and methods for processing and analyzing these data. The Heads of Agencies and the European Medicines Agency Joint Big Data Taskforce was established to consider these issues from the regulatory perspective. This review reflects the thinking from its first phase and describes the big data landscape from a regulatory perspective and the challenges to be addressed in order that regulators can know when and how to have confidence in the evidence generated from big datasets.**

The role of medicines regulatory agencies is multifactorial and broad; their overarching responsibility is to ensure that all approved medicines, medical devices, and the combination thereof, are both effective and safe but in order to deliver on that responsibility, medicines regulators must work across the entire drug development pathway. For example, there is a need not only for evidence on the intended and unintended effects of medicinal products arising from both the highly controlled environment of a randomized controlled clinical trial and subsequently in clinical practice but also for information on disease, its prevalence, and progression across a population, on current standards of care across our diverse European population and on prevalence of potential adverse effects to contextualize information around medicinal products. Increasingly there is a need for an understanding of the accuracy of diagnostic tests, including imaging tests, which may impact on either the diagnosis of a disease and, hence, the prescribing of a medicinal product or the monitoring of its effectiveness and/or safety.

The unparalleled pace of change in the scientific landscape is challenging the current regulatory paradigm and requiring 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 the analytic methods used and, hence, the veracity of the evidence ultimately generated.

Although the term *big data* is widely utilized, there is no commonly accepted definition and the concept is quite nebulous with no universally defined thresholds for any of the presumed characteristics. In our view, a definition should encompass not only the concept that big data is diverse, heterogeneous, and large, and incorporates multiple data types but should also refer to the complexity and challenges of integrating the data to enable a combined analysis. Hence, we define big data as extremely large datasets, which may be complex, multidimensional, (un)structured, and heterogeneous, which are accumulating rapidly and which may be analyzed computationally to reveal patterns, trends, and associations. In general, big datasets require advanced or specialized methods to provide an answer within reliable constraints. Thus, a single dataset may not strictly meet the definition of big data but when pooled with other datasets of a similar type, or linked to other datasets of different types, the datasets become sufficiently large or the difficulties in pooling, linking, and analyzing are sufficiently complex for the data to assume the characteristics of big data.

Datasets of most immediate utility for regulatory decision making are data derived from previous clinical trials, real-world data, including postmarketing registry data, spontaneous adverse drug reaction (ADR) reports, and genomic data, especially if linked to clinical data. Other types of 'omic data, such as proteomics and metabolomics, represent more heterogeneous data and at the far end of the uncertainty scale, sits individually generated social media data. There is little doubt that the development of future treatments (medicines, *in vitro* diagnostics, devices, or digital therapeutics) will utilize such data, which may reach regulatory authorities either as supportive data together with more traditionally analyzed structured data, or may underpin the submission as a whole. Thus, it is essential that regulators understand its presence and the robustness by which it was generated in order to make a competent evaluation of the submission as a whole and to continuously monitor the medicine, device, or the performance of the *in vitro* diagnostics on the market.

This challenge is significant: The paradigm for authorization in most stringent regulatory authorities is based on the assessment of well-controlled, randomized, high-quality data of known

[1]European Medicines Agency, Amsterdam, Netherlands; [2]Danish Medicines Agency, Copenhagen, Denmark. *Correspondence: Alison Cave (alisoncave@hotmail.co.uk)

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agencies or organizations with which the authors are employed/affiliated.

**Received October 28, 2019; accepted December 12, 2019. doi:10.1002/cpt.1736**

provenance. In contrast, big data offers evidence that may be derived from unstructured, heterogeneous, and unvalidated data of potentially unknown provenance, often with unknowns around potential bias and with additional uncertainties of accuracy and precision. Much of this information may be at the individual level. Moreover, not all datasets are the same; there is variable quality and standardization, data are generated under different scenarios and for different purposes, and ownership resides with multiple stakeholders, many of whom have no obligation to engage with regulatory systems. Thus, influencing the data landscape to meet regulatory needs is complex. Furthermore, with the availability of multiple linkable datasets, many associations will be observed, which may or may not be spurious, but which will consume considerable resources to validate and may generate significant concerns. Processes will need to be defined to replicate and test findings and also to determine how to define the threshold of evidence required for regulators to act.

Generalizing that evidence threshold is difficult when evidence suitable for regulatory decision making can be a variable concept. As such lower grades of evidence may be acceptable in the case of rare diseases where data are hard to collect compared with more frequent diseases, or when assessing a safety risk where randomized evidence is impossible to obtain. Nonetheless, certain evidentiary standards have evolved as scientific methodology and rigor has advanced from pure empirical thinking relying on serial observations as the basis for causal inferences through Popper's falsification principles on to present day's randomized double-blind placebo-controlled clinical trials as the gold standard for evidence. When considering real-world data for use in a regulatory context the point where data becomes evidence is of crucial importance. A large dataset with a large series of observations does not in itself automatically constitute good evidence of causality but when and why evidence reaches the causality threshold needs to be clarified. Regulators must set the standard for good evidence and guide innovators and industry in their efforts to generate evidence suitable for regulatory decision making.

Given this context, the task force of the EU Heads of Agencies and the European Medicines Agency (HMA-EMA Joint Big Data taskforce) was formed to describe the big data landscape from a regulatory perspective in order to inform the EU regulatory network on decisions and planning on the capability and capacity to guide, analyze, and interpret these data. This paper is a distillation of the report from the first phase of the taskforce and sets out the thinking of the Task Force around the steps that need to be taken in the wider data landscape in order that regulators can know when and how to have confidence in the evidence generated from big datasets. A fuller version of this report and the reports of the individual subgroups are available on the EMA and HMA websites. As for many stakeholders, regulatory concerns center not only around the data itself, its quality (including any data transformations), and representativeness, but also on the analytical processes (data manipulation, modeling, and analytics) used to generate the evidence. It is clear that delivery of the Task Force recommendations will require the consolidated action of multiple stakeholders and substantial resources. The Task Force has continued its work and its phase II report, proposing the priorities, timing, and approach to implementing recommendations for big data in medicines regulation, is anticipated for publication in 2020.

## DATA CHARACTERISTICS

To a large extent, data quality determines the validity of the evidence that can be reliably derived from a given dataset. Thus, understanding quality is a key need. However, quality is hard to define and it is even harder to prespecify what the required data quality attributes might be over a range of regulatory use cases. Acceptability will always be influenced by the context of use; the opportunities and timeliness of other data capture options, the question being asked, the level of risk associated with each decision, the availability of other treatments, and the unmet medical need (for recent examples see ref. 1). Thus, guidelines are needed to inform on regulatory expectations around data quality across the range of regulatory decisions in the context of the risk associated with each decision, which should, where possible, seek alignment with other regulatory authorities.

To move the dial, we require the capability to characterize data based on a data quality framework, which enables a common understanding of the strengths and limitations of big datasets. Ideally, such a framework should be model and geography agnostic to allow a comparison across data sources of the same type whatever their origin. Moreover, the framework should address more than just the data content, but also the quality control measures in place to describe reproducibility over time, and these findings should be transparently recorded in a sustainable, accessible inventory. Standards should include measurement technologies to allow systematic benchmarking and validation as these analytical techniques evolve with a particular focus on the reproducibility of results. The 2009 report by the Human Proteomics Organization test Sample Working Group illustrates the scale of the challenge: of 27 laboratories examining the same sample that consisted of 20 highly purified proteins, only 7 laboratories reported all 20 proteins correctly.[2] Thus, achieving sustainability of such a venture will require collaboration, engagement, and sustained effort over multiple stakeholders to maximize its utility. On a much smaller scale, the current EMA Patient Disease Registry Initiative provides an example of how incorporating the needs of all relevant stakeholders informs the development of minimal quality standards and data elements in order to facilitate downstream data harmonization and maximize the utility of the data.[3] Facilitating broad engagement and seeking agreement will support and drive adoption.

A key enabler for data characterization is standardization in order to drive harmonization across datasets, enhance interoperability, improve data quality, allow comparability, and facilitate data analyses. However, standardizing data is hard; much of the data are unstructured and heterogeneous, and this is especially true of social media data and data from wearables that are anticipated to account for much of the data volume increases in the coming years.

The need for standards is not new. It was recognized many years ago, and when required for regulatory purposes has driven global harmonization, a key component of the mission of the International Council for Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use. For instance,

the data model and data elements of the Individual Case Study Report, which is used for reporting of ADRs, provides clear specification of reporting requirements for ADRs.[4] As a result, platforms, such as EudraVigilance,[5] contain extremely well-structured information, although the completeness and accuracy of the information within Individual Case Study Report forms is still dependent on the reporter and, hence, variable. However, many other datasets are not standardized partly because most have evolved over many years, and, hence, the data encompass many technological developments and partly because there is not one single owner. In addition, with the exception of 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 worldwide, but while relatively well structured, much of the data is siloed by disease, institution, and country, generated with different methodologies, analyzed by nonstandardized software, and often stored in incompatible file formats, and consequently only a small percentage is linked.[6] This situation is replicated over multiple datasets in the big data landscape. Consider also that genomics is one of the most well organized fields with considerable harmonization efforts, such as Global Alliance for Genomics and Health Connect, already underway.[7] The resources to duplicate such efforts across the full spectrum of datasets will be considerable.

No single data standard will have the depth and breadth to be applicable to all datasets. However, as a community, we should strive as much as possible to minimize the number of standards. To encourage adoption, standards should be transparent and open source to promote widespread uptake, globally applicable, comprehensive, and maintained with an ongoing process for testing and revision that is sustainable. It is, therefore, important to strongly support the use and maintenance of available data standards, and the development of standards where none are available (e.g., novel data sources, such as m-health, and less mature fields, such as epigenetics, to ensure early alignment).[8] A recent example of regulatory uptake and support of standards is provided by International Organization for Standardization (ISO) Identification of Medicinal Products,[9] which aims to facilitate international identification of medicinal products.

Although the benefits of standardization are clear, implementation of standards is expensive, and the return on investment must be sufficient to outweigh these costs. The cost of implementation of ISO Identification of Medicinal Products alone across 14 European Federation of Pharmaceutical Industries and Associations member companies has been estimated to be at least ~ €70M.[10] Hence, 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. Sustainability of standards is challenging but will be enabled by widespread adoption. From a regulatory perspective, a prioritization of efforts will be needed with an early focus on data most likely to impact on decision making in the near term.

At a global level, it is important to ensure that extremely expensive and time-consuming standardization and data mapping initiatives[11–13] do not pull in opposite directions but work together to achieve sustainable and global solutions. From a regulatory perspective, global cooperation is important, as for many rare diseases and cancers or indeed rare ADRs there may only be a handful of cases worldwide, and these data need 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, standardization 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).

Standardization will not only enable better data characterization but will also facilitate data linkage between related datasets to provide additional insight not possible from single isolated datasets. This is a key requirement as European healthcare data are heterogeneous; differences in healthcare systems, national guidelines, and clinical practice have driven different content, and, hence, the generalizability of a single healthcare system from a single European country cannot be assumed for the whole of Europe. Moreover, data linkage applies not only to databases within a subgroup (e.g., how to integrate different registries or electronic health records but among disparate datasets; e.g., linking clinical data with genomic/pharmacogenomics data and proteomic data and linking data across care settings; e.g., primary, secondary, and tertiary care). For example, predicting a patient's response to a therapeutic intervention with a proteomic or genomic biomarker in order to minimize exposure of patients to ineffective or intolerable therapies, can only be achieved if 'omics data is linked to clinical outcomes. Unfortunately, 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), thereby limiting their value in a regulatory context. Raising awareness of the need for linkage of treatment and outcome data would be particularly beneficial.

Different questions will require linkage of data at different levels and require different data protection solutions. For some regulatory needs, linkage at an individual patient level would ideally be required (e.g., understanding the clinical outcome of an ADR or enabling longitudinal follow-up of a genomic targeted or gene-editing medicine). However, there are many scenarios where linkage of data at a population level would be sufficient (e.g., standard of care at different disease stages across Europe or outcomes from vaccination programs). To enable meaningful data linkage, sharing needs to move beyond simply sharing the raw data, to encompass associated metadata, which describes key characteristics about the data (e.g., sample type, disease stage, treatment, and genomic mutation).[14]

## DATA ANALYTICS

Big Data analytics is a growing field of data science, which combines methods from various disciplines, including biostatistics, mathematical modeling and simulation, bio-informatics, and computer science, and encompasses data collection, data management, data-integration, data standardization, machine learning (ML), and requires specialized information technology

architectures and tools to extract knowledge and insights from data in its various forms both structured and unstructured. It is, however, essential to remember that data by itself does not provide value: It needs to be analyzed, interpreted, and acted upon. Hence, methodologies and analytical approaches are equal partners in generating evidence from big data.

Artificial intelligence (AI), defined as a self-learning evolution of well-known, advanced adaptive statistics (for review see ref. 15), is gaining much attention as a route to extract the greatest value from big data and is already part of the regulatory landscape. ML algorithms, where the algorithm incorporates feedback to continuously optimize the output, have been incorporated in randomization algorithms and many devices/apps can or may use these techniques.[16] Further development of AI into natural language processing (NLP) recognizing and processing free text,[17] multilayered perceptron algorithms, recognizing images,[18] robotics guiding surgical instruments, and even deep learning (DL) algorithms,[19] are now part of the data handling landscape regulators must consider.

AI has scientific utility in three main areas: Descriptive, in terms of providing a quantitative summary of selected features of the data; prediction, where a number of inputs are mapped to an output within the data in order to predict a future, unobserved event; and causal inference to allow conclusions to be drawn about a causal association between an occurrence and an effect. In terms of the regulatory context, three immediate areas of AI applicability seem urgent to address:

1. Regulatory approval of AI-based health apps in devices intended for clinical decision making.
2. Regulatory evaluation of AI derived evidence, predominantly from ML, on the effectiveness and safety of medicinal products. In such scenarios, complex algorithms may have been involved in the identification and matching of patients for inclusion in clinical trials, in the generation of complex outcome measures, or may have been integral to the processing or manipulation of such data. One additional obvious area of application is in the analysis of spontaneous ADRs where advanced analytical solutions, which take advantage of the richness of information available in the regulatory databases, but can also exploit information in other databases to describe patterns and associations, could deliver a faster and more accurate identification of safety signals.[20] Equally, NLP will have utility in extracting structured information from case narratives associated with spontaneous ADRs and clinical notes linked to electronic health records to support decision making.
3. Regulatory use of AI in internal processes to increase efficiency. For example, NLP processing of text, categorizing electronic common technical document submissions into review templates for assessors or quantitative multilayered perceptron review of image data submitted to support a clinical claim from a drug manufacturer.

Fundamental to the role of the regulator is clearly the assessment of the benefit-risk balance of a medicinal product at authorization and over time, which usually brings the need to understand the causality of an observed association. With the increasing digitization of healthcare data, there is significant interest in the use of large observational health data sources to complement and support regulatory decision making across the product life cycle. In this case, the acceptability of evidence generated from this real-world data is not only dependent on data quality, as discussed in earlier sections, but also on the methodological processes used to generate the evidence and on the measures implemented to control for bias and confounding; this is especially true for evidence generation from observational data sources in the absence of randomization and is key for understanding whether any association is causal or merely random. Increasingly, data scientists are using automated data adaptive techniques to optimize the control of confounding,[21,22] which importantly in the context of the heterogeneous European data landscape, claims to be dataset agnostic and, thus, applicable to any structured or unstructured data source and coding system. Although automated approaches bring advantages, it could also be argued that the loss of human investment in the building of the model (from creating and testing the models where many choices require intuition from analysts and subject matter experts, to working with the model results) is a significant disadvantage. While the best approach is still the subject of debate,[23] improving confidence in the control of confounding would undoubtedly significantly increase confidence in the causality of any association arising from the application of an ML algorithm.

From a regulatory perspective it is clear that it is the validation of these novel analytical approaches in order to understand the associated limitations and risks, which will be a key part in defining their acceptability. Multiple challenges become apparent when one considers the dynamic, constantly evolving the nature of ML (not least DL where the algorithm is optimizing itself toward better performance), and its application to regulatory questions. A synopsis of the main issues is presented here but an in-depth review of the impact of AI in the regulatory setting is available in the data analytics report of the HMA-EMA Joint Big Data taskforce.[24]

The performance of an algorithm is dependent on the datasets it is trained on. Thus, the output will reflect the distribution, variability, and complexity of the data in the training dataset and potentially the bias of those training the algorithm. This has implications for the generalizability of the algorithm beyond the data used to train and fit the model. For example, if an algorithm is trained on predominantly Western European data, it may not be predictive of outcomes for Southern/Eastern European populations or immigrants to Europe of African descent. What then is the uncertainty of the algorithm in predicting the outcome of interest when applied across the entire European population?

Second, it will also be important to understand what performance metrics the algorithm was trained upon. For example, was the sensitivity of the algorithm the predominant driver in evaluating performance or its accuracy? The US Food and Drug Administration (FDA) has recently approved ML-based algorithms for the diagnosis of diabetic retinopathy[25] and detection of wrist fractures,[26] and the consequences of a misdiagnosis or delayed diagnosis are all too apparent. Patient safety clearly demands sufficient regulation. It is, therefore, imperative that regulators require algorithms to "explain themselves" (i.e., to be programmed in advance with a view toward interrogability and interpretability). Additionally, ML algorithms need to flag data where predictions outside of the

distribution of the training dataset may not be accurate or reproducible. It may be difficult to use classical validation approaches for ML technology but in several areas, it will be possible to validate against gold standards (e.g., in quantitative imaging analysis where measurements done by trained radiologists on Digital Imaging and Communications in Medicine-standard images can serve as the standard). At least initially, DL algorithms pose a separate challenge, as they may not be readily open for interrogation/validation.

By their very nature, ML algorithms are in constant change and evolution, and the difficulty of how, and especially when, to evaluate the outcomes becomes apparent. 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. Clearly, the network needs to define the acceptable circumstances of outcomes based upon such technology and this will be challenging across multiple use cases. One approach could be that it is accepted that by their very nature one cannot validate them, but one can still try and falsify their outcome (e.g., by programming other algorithms to attempt falsification of their very premises). Only if these attempts fail can one accept the outcome of the primary algorithm (i.e., an evolution of Karl Poppers general falsification theory).

Big data has a reproducibility challenge not only because the datasets are dynamic with sometimes unknown provenance, but because metadata is not always fully described, which makes it very challenging to document the data and analytical journey. Transparency will be key in delivering trust and, as such, data platforms, which incorporate mechanisms to increase the transparency of the data and the analysis, are to be encouraged. Options include bioinformatics applications addressing metadata documentation, standardization, annotation and data management, open source, user-friendly algorithms and tools, and direct coupling to dedicated and performant statistical analysis. Agreement from stakeholders to describe their data in a comprehensive and standardized manner will significantly increase replicability but requires constant engagement with all relevant factors.[14] Utilization of existing regulatory processes, such as the EMA Qualification Advice, will enable regulators to influence more mature approaches, and we see increasing interest in this process.[27–29] It is perhaps in understanding the comparability of outcome measures produced by such approaches that the biggest challenges will be faced. Discussions in such fora would be significantly enhanced by a framework to support regulators in performing a systematic and consistent evaluation of ML algorithms across applications. As such, the FDA has recently posted a discussion paper on Good Machine Learning practice[30]; similar discussions need to be progressed in Europe. Last, 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 NLP techniques.

There is no doubt that AI has the potential to greatly improve data handling, processing, and even accuracy and predictability in health care. However, it must be held accountable to regulatory standards, just like the framework for pharmacovigilance was established in the wake of the Thalidomide scandal in the 1960s. It is clear that regulators cannot and should not accept the so-called "black box" concept where algorithms simply perform in a vacuum without any checks and balances. Algorithm code should be available for review by regulators, and outcomes of algorithm use (safety and efficacy) needs to be subject to postmarketing surveillance mechanisms, just like is done today to monitor drug safety after marketing authorization. Moreover, it is imperative that regulators and decision makers now invest in upskilling their staff and the regulatory infrastructure to meet these new challenges. Only then can the true potential of these technologies be safely deployed.

## DATA SHARING AND ACCESS

Fundamental to any big data vision is the need to share and access data in a timely fashion. 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.

Data sharing is motivated by the belief that sharing and integrating data across multiple datasets maximizes 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 increasingly demanding that data generators, be they academics, healthcare professionals, or industry, commit to meaningful data sharing practices (e.g., EMA Policy 0070 (EMA/240810/2013)).[31]

Despite the recognized benefits of data sharing, multiple barriers are preventing its natural progression, some of which are common across datasets.[32] First, it is becoming progressively more challenging to share increasingly complex data from multiple sources in sufficient depth and detail so as to retain its scientific utility and meet data protection obligations on a global scale. Robust data anonymization 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 anonymization are urgently needed to resolve this dilemma and find an appropriate balance and consistency of approach to derive the benefits of data sharing. Clearly, patients must be the partners in these discussions and in the development of such principles.

It is recognized 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, are a critical part of any study. In addition, early consideration ensures that the budgetary planning for resources required to make data accessible is considered at the inception of projects and built into onward sustainability plans.

To derive maximum benefit, data needs to be shared at a sufficient level of detail. Data sharing platforms should mandate sharing of metadata, and as a prerequisite 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 harmonization and pooling of datasets. It is notable that Europe has failed to define a clear path to enable the sustainability of many previous data sharing efforts, particularly for observational healthcare data, and defining this should be a priority in the future.[33] 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 coordinated mechanism for funding infrastructure platforms across Europe may allow the provision of continued funding for those platforms that 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, funder initiatives, such as the Horizon 2020 Open Research Data pilot,[34] and measures from journals to share data underlying published papers.[35,36] However, additional policy initiatives are needed to truly promote a data sharing culture that is mutually beneficial for, and applicable to, all stakeholders. In the commercial sector, a recent analysis suggests that policies on the sharing of trial data, results, and methods across pharmaceutical companies is highly variable[37] and, hence, measures to increase transparency around data sharing in this sector are still urgently needed especially where trials did not form part of a European regulatory application and, therefore, fall outside of Policy 0070. For academics, 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 opinions). 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 should be shared (with appropriate data protection measures) to enable linkage to the disease and clinical outcome data with which they are associated.

Much of the promise of big data requires the ability to link and interrogate multiple different types of data. Although a worthy goal, it is appreciated that 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,[38,39] block chain and homomorphic encryption, and which meet national and international data protection legalization are urgently required. Distributed datasets where personal identifiable data are 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 and seems the most realistic and feasible solution to enable data accessibility. Recent initiatives, such as the Beacons project,[40] provide mechanisms not only for data discoverability but also for onward data sharing. It utilizes a simple application program interface, which once implemented, allows users to query the existence of specific information. In the case of the Beacons network, a database can then choose to share more data around the specific request and moreover can choose the level of that data disclosure. Such probe technology can be implemented more widely as the ability to probe a dataset or indeed multiple datasets in parallel, run an algorithm, and return an anonymized answer at an aggregate level will not only increase data discoverability but also overcome privacy concerns around the sharing of patient level data. Although distributed/federated data models do have the drawback of losing some statistical power compared with a model with one common data repository, even small degrees of data optimization (standardization of the distributed data toward a common set of accepted standards) might ameliorate this.

## CONCLUSIONS

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 authorization, and which, in turn, must be carefully managed postauthorization. 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, predefined 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 utilization. The actions outlined in this review and summarized in **Table 1**, particularly increased standardization 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 for that guidance 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

**Table 1 Summary of recommendations from phase I of the HMA-EMA Joint Big Data Task Force**

| | |
|---|---|
| Promote use of global, harmonized, and comprehensive standards to facilitate interoperability of data | • Minimize 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.<br>• Where data cannot be standardized at inception, establish the regulatory requirements to confirm the validity of mapped data.<br>• Promote use of global open source file formats. |
| Characterization of data quality across multiple data sources is essential to understand the reliability of the derived evidence | • Characterize and document data quality in a sustainable EU inventory.<br>• 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 datasets for analysis).<br>• Implement data quality control measures.<br>• Establish a clear framework for the validation of innovative bioanalytical methods (e.g., 'omics). |
| The development of timely, efficient and sustainable frameworks for data sharing and access is required<br>Further support mechanisms are needed to promote a data sharing culture | • Strongly recommend the establishment of distributed data networks to facilitate data sharing of sensitive healthcare data.<br>• Develop guidance for robust data governance and data anonymization to deliver systems that secure patient trust.<br>• Establish disease-specific minimum data elements to enable harmonization of data across, for example, national disease registries.<br>• Promote mandatory sharing of the analysis arising from data sharing activities (e.g., by publication or open sharing via data access platforms).<br>• Promote the sharing of qualified models.<br>• Support the development of policy initiatives to drive a data sharing culture, which is mutually beneficial for all stakeholders. Patients should be partners in all discussions.<br>• Proactively drive and/or support data sharing platforms and initiatives.<br>• Require the submission of data management plans at the start of all data generation exercises.<br>• Establish accountability for users.<br>• Development of common principles for data anonymization to facilitate data sharing. |
| Promote mechanisms to enable data linkage to deliver novel insights<br>Facilitate harmonization of similar datasets | • Encourage sharing of raw data, associated metadata and processed data to enable meaningful data linkage.<br>• Proactively engage with initiatives to map terminologies to facilitate data linkage and timely data access but ensure frameworks for consistent validation are simultaneously implemented.<br>• Support mechanisms to maintain up-to-date mappings across terminologies.<br>• Promote the inclusion of clinical outcome data relevant to regulatory questions in public databases. |
| Develop clear frameworks to enable the validation of analytical approaches to determine if they are appropriate to support regulatory decision making<br>Promote new analytical approaches for modeling of big data sets for regulatory purpose | • 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.<br>• Form an advisory group to:<br>◦ explore the applicability of novel analytics methodologies to support the development, scientific evaluation, and monitoring of medicinal products;<br>◦ explore the most suitable data standards and IT architecture and tools capable to enable the analyses.<br>• Promote the increased utilization of scientific advice and the EMA Qualification Advice process to enable regulators to influence more mature approaches.<br>• Support, define, and validate the definition of innovative outcome measures and other approaches, which leverage additional dimensions from high-frequency or high-dimensional data.<br>• Explore novel methodologies to improve the control of confounding in observational studies and other big data studies.<br>• Make publicly available data analysis plans for all studies submitted for regulatory approval.<br>• Strongly support the exploration of novel analytics approaches, such as natural language processing techniques to interrogate unstructured data.<br>• Agree and create guidelines on which level of validation, reproducibility, and trustworthiness of evidence is acceptable according to the regulatory application of the AI algorithm. |
| Regulatory guidance is required on the acceptability of evidence derived from big data sources | • Identify the best format to enhance the agility of guidance development and revision in this fast moving field.<br>• Track concrete examples of procedures relevant to big data across the regulatory network to inform thinking.<br>• Establish pilot programs to develop informal discussion on acceptability.<br>• Initiate pilot studies to better understand the evidence generated on efficacy/effectiveness and safety from emerging datasets.<br>• Mandate transparency and format around study reporting for regulatory submission to document datasets, protocol, tools, and version used to promote reproducibility.<br>• Emphasize the need for outcome measures from novel data sources (e.g., m-health devices to be reflective of a defined clinical benefit). |

Full details of the recommendations can be found in the phase I summary report of the HMA-EMA Joint Big Data Taskforce.[46]
AI, artificial intelligence; EU, European Union; GCP, good clinical practice; HMA-EMA, Heads of Agencies and the European Medicines Agency; IT, information technology.

in place to guarantee the data elements. For digitally captured data, quality measures would need to incorporate the algorithms used 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 end points provided by wearables for utilization in clinical trials,[41] ingestible sensor systems for adherence,[42] on novel biomarkers,[43] and on data sources appropriate for regulatory decision making.[44,45] 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 can 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.

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, also 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. 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. 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 the scientific community and regulators 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 analyze, interpret, and profit from the data generated. In this way, we will improve our decision making and enhance our evidentiary standards.

**ACKNOWLEDGMENTS**

The contribution of all members of the taskforce is gratefully acknowledged. The full membership is listed below in alphabetical order: Ada Georgescu, RO; Aldana Rosso, DK; Alexandra Pacurariu, EMA; Alison Cave, EMA; Antti H. Hyvärinen, FI; César Hernandez Garcia, ES; Didier Meulendijks, NL; Didier Meulendijks, NL; Dieter Deforce, BE; Gavril Flores, MT; Gianmario Candore, EMA; Hans Ovelgönne, NL; Katherine Donegan, UK; Kevin Horan, IE; Luis Pinheiro, EMA; Marek Lehmann, EMA; Marjon Pasmooij, NL; Mark Goldammer, DE; Martin Nyeland, DK; Mateja Sajovic, SI; Miguel Ángel Maciá, ES; Nikolai Brun DK; Panagiotis Telonis, EMA; Paolo Alcini, EMA; Per Fuglerud, NO; Renate König, DE; Roxana Dondera, RO; Roxana Stroe, RO; Thomas Senderovitz, DK; Vesa Kiviniemi, FI; Zsuzsanna Cserjes Szabone, HU; Zsuzsanna Szabóné Cserjés, HU. We would also like to acknowledge the support of Peter Arlett (EMA) in the finalization of the phase I report of the Task Force and for his comments on this manuscript. In addition, we extend our thanks to Jolanta Palepsaitiene for secretarial assistance with the manuscript.

**FUNDING**

No funding was received for this work.

**CONFLICT OF INTEREST**

The authors declared no competing interests for this work.



1. Cave, A., Kurz, X. & Arlett, P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin. Pharmacol. Ther.* **106**, 36–39 (2019).
2. Bell, A. *et al*. HUPO Test Sample Working Group. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**, 423–430 (2009).
3. McGettigan, P. *et al*. Patient registries: an underused resource for medicines evaluation: operational proposals for increasing the use of patient registries in regulatory assessment. *Drug Saf.* **42**, 1343–1351 (2019).
4. Santoro, A., Genov, G., Spooner, A., Raine, J. & Arlett, P. Promoting and protecting public health: how the European Union pharmacovigilance system works. *Drug Saf.* **40**, 855–869 (2017).
5. EudraVigilance.EudraVigilance is the system for managing and analysing information on suspected adverse reactions to medicines which have been authorised or being studied in clinical trials in the European Economic Area (EEA) <http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000679.jsp>.
6. Stark, Z. *et al*. Integrating genomics into healthcare: a global responsibility. *Am. J. Hum. Genet.* **104**, 13–20 (2019).
7. GA4GH Connect: a 5-year strategic plan <https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf>.
8. Marti-Renom, M.A. *et al*. Challenges and guidelines toward 4D nucleome data and model standards. *Nat. Genet.* **50**, 1352 (2018).
9. Substance, product, organisation and referential (SPOR) master data <https://www.ema.europa.eu/human-regulatory/research-development/data-medicines-iso-idmp-standards/substance-product-organisation-referential-spor-master-data>.
10. Principles for the Implementation of ISO IDMP Standards for EudraVigilance and Development of a Road Map <https://www.efpia.eu/media/25717/principles-for-the-implementation-of-iso-idmp-standards-for-eudravigilance-and-development-of-a-road-map-2014.pdf>.
11. European Open Science Cloud (EOSC) <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>.
12. The Yosemite Project <http://yosemiteproject.org/>.
13. EHDEN <https://www.ehden.eu>.
14. Wilkinson, M.D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
15. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).
16. Dart, A. The algorithm will see you now. *Nat. Rev. Cancer* **17**, 42 (2017).
17. Banerjee, I., Bozkurt, S., Caswell-Jin, J.L., Kurian, A.W. & Rubin, D.L. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO Clin. Cancer Inform.* (2019). https://doi.org/10.1200/CCI.19.00034. [e-pub ahead of print].

18. Gulshan, V. *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
19. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
20. Basile, A.O., Yahi, A. & Tatonetti, N.P. Artificial intelligence for drug toxicity and safety. *Trends Pharmacol. Sci.* **40**, 624–635 (2019).
21. Schneeweiss, S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin. Epidemiol.* **10**, 771–788 (2018).
22. Sofrygin, O. *et al*. Targeted learning with daily EHR data. *Stat. Med.* **38**, 3073–3090 (2019).
23. Tian, Y., Scheumie, M. & Suchard, M.A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int. J. Epidemiol*. **47**, 2005–2014 (2018).
24. Data Analytics subgroup report, November 2019 <https://www.hma.eu/fileadmin/dateien/HMA_joint/00-_About_HMA/03-Working_Groups/Big_Data/2019_11_HMA-EMA_Big_Data_TF_Data_analytics_subgroup_report.pdf>.
25. US Food and Drug Administration <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm>.
26. US Food and Drug Administration <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm608833.htm>.
27. Manolis, E., Vamvakas, S. & Isaac, M. New pathway for qualification of novel methodologies in the European Medicines Agency. *Proteom. Clin. Appl.* **5**, 248–255 (2011).
28. Manolis, E., Koch, A., Deforce, D. & Vamvakas, S. The European Medicines Agency experience with biomarker qualification. *Methods Mol. Biol.* **1243**, 255–272 (2015).
29. Isaac, M., Vamvakas, S., Abadie, E., Jonsson, B., Gispen, C. & Pani, L. Qualification opinion of novel methodologies in the predementia stage of Alzheimer's disease: cerebro-spinal-fluid related biomarkers for drugs affecting amyloid burden-regulatory considerations by EMA focusing in improving benefit/risk. *Eur. Neuropsychopharmacol.* **21**, 781–788 (2011).
30. US Food and Drug Administration <https://www.fda.gov/media/122535/download>.
31. Taichman, D.B. *et al*. Data sharing statements of clinical trials: a requirement of the International Committee of Medical Journal Editors. *Ann. Int. Med.* **167**, 63–65 (2017).
32. Bierer, B.E., Crosas, M. & Pierce, H.H. Data authorship as an incentive to data sharing. *N. Engl. J. Med.* **377**, 402 (2017).
33. Plueschke, K., McGettigan, P., Pacurariu, A., Kurz, X. & Cave, A. EU-funded initiatives for real world evidence: descriptive analysis of their characteristics and relevance for regulatory decision-making. *BMJ Open* **8**, e021864 (2018).
34. European Commission. H2020 online manual <http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm>.
35. International Committee of Medical Journal Editors. Journals stating that they follow the ICMJE recommendations <http://www.icmje.org/journals-following-the-icmje-recommendations/>.
36. Wiley. Data Sharing Policy <https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html>.
37. Goldacre, B., Mahtani, K.R., Heneghan, C., Onakpoya, I., Bushfield, I. & Smeeth, L. Pharmaceutical companies' policies on access to trial data, results, and methods: audit study. *BMJ* **358**, j3334 (2017).
38. Gaye, A. *et al*. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int. J. Epidemiol.* **43**, 1929–1944 (2014).
39. Wilson, R.C. *et al*. DataSHIELD – new directions and dimensions. *Data Sci. J.* **16**, 21 (2017).
40. Fiume, M. *et al*. Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).
41. European Medicines Agency. Draft qualification opinion on stride velocity 95th centile 5 as a secondary endpoint in Duchenne Muscular Dystrophy 6 measured by a valid and suitable wearable device (EMA/CHMP/SAWP/527447/2018) <https://www.ema.europa.eu/documents/regulatory-procedural-guideline/draft-qualification-opinion-stride-velocity-95th-centile-secondary-endpoint-duchenne-muscular_en.pdf>.
42. European Medicines Agency. Qualification opinion on ingestible sensor system for medication adherence as biomarker for measuring patient adherence to medication in clinical trials (EMA/CHMP/SAWP/513571/2015) <https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-ingestible-sensor-system-medication-adherence-biomarker-measuring-patient_en.pdf>.
43. European Medicines Agency. Qualification opinion on plasma fibrinogen as a prognostic biomarker (Drug Development Tool) for all-cause mortality and COPD exacerbations in COPD subjects (EMA/CHMP/SAWP/264260/2018) <https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-plasma-fibrinogen-prognostic-biomarker-drug-development-tool-all-cause_en.pdf>.
44. European Medicines Agency. Qualification Opinion on The European Cystic Fibrosis Society Patient Registry (ECFSPR) and CF Pharmaco-epidemiology Studies (EMA/CHMP/SAWP/622564/2018) <https://www.ema.europa.eu/documents/regulatory-procedural-guideline/qualification-opinion-european-cystic-fibrosis-society-patient-registry-ecfspr-cf-pharmaco_en.pdf>.
45. European Medicines Agency. Draft qualification opinion on cellular therapy module of the European Society for Blood & Marrow Transplantation (EBMT) Registry (EMA/CHMP/SAWP/423488/2018) <https://www.ema.europa.eu/documents/regulatory-procedural-guideline/draft-qualification-opinion-cellular-therapy-module-european-society-blood-marrow-transplantation_en.pdf>.
46. HMA-EMA Big data taskforce. Summary report <https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf>.