EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

Identifying opportunities for 'big data' in medicines development and regulatory science

Report from a workshop held by EMA on 14–15 November 2016

Summary

Recent rapid expansions of health and biological data, combined with the availability of sophisticated computational technologies, offer unprecedented opportunities to benefit public health. It is timely that the European regulatory network has, with this workshop, sought to define the landscape of the big data field in order to clarify the possibilities and challenges and identify how medicines regulators can exploit big data to support medicines development and regulatory decision-making.

In the medicines field, expanding data sets from, for example, electronic health records and mobile devices, and new insights gleaned from analysis of entire organisms and biological systems, offer tantalising potential to benefit patients and consumers. While collection and use of data come with the responsibility to ensure security and protect patient privacy, patient groups expressed their willingness to share data in the knowledge that it advances research and ultimately helps other patients.

Massive and continually increasing amounts of healthcare data are being produced in all dimensions: data from the population level down to the individual, health and disease data, diverse data types from traditional health measurements to 'omics and social media, and information on patients' experiences through time. Unfortunately, most of these data are unstructured and stored in poorly curated repositories or in silos, and thus inaccessible. To maximise the opportunities for public health, data must be made findable, accessible, interoperable and reusable (FAIR). Integration of data from various sources for subsequent analysis is a further challenge. However this may be overcome using approaches such as common data models: the Observational Medical Outcome Partnership (OMOP) and the Sentinel Common Data Models were presented during the workshop.

Once data are available and in a structured format, computational technologies for analysis,

such as machine learning and data mining, already exist. Many insights from big data analysis were presented during the workshop including examples in target discovery, drug-drug interactions, image analysis, mapping vaccine uptake, patterns of medicine use and prediction of disease. Although storage and analysis of huge data sets is computationally intensive, use of the cloud instead of on-site systems can help provide the necessary capacity.

However, big data is noisy and there may be missing data, and known or unknown biases. It can be difficult to create trust in results and to establish causality and rule out coincidence, factors that are essential for evidence in medicines regulation. Approaches to overcome these challenges discussed during the workshop include analysing lower quality data to create hypotheses, followed by corroboration with careful retrospective analysis of higher quality, more meticulously selected data and laboratory experiments; performing randomised trials with real world data; and using data sources to supplement prospective randomised clinical trials.

In addition, to avoid bad analysis of observational data and increase robustness of results, the health and research community should come together to agree best practices, develop open-source analytical tools and establish quality standards for observational data analogous to those currently in place for clinical trials.

Rather than the development of proprietary tools and restriction of access to data, an 'open science' approach to big data, incorporating public-private collaborations, was favoured by workshop participants. This would create networks of people implementing new technologies, approaches and standards and promote a collaborative environment driving sharing of knowledge.

The impact of both big data and real world data (data collected outside randomised clinical trials usually during normal clinical care) is significant across the medicines life cycle. There are many existing and potential applications: in drug discovery, promising compounds for screening can be selected in silico; in clinical trials, epidemiological information can inform trial designs in terms of defining outcome measures and site selection for optimal recruitment; in the post-marketing stage, real world data can be used to continually assess the benefit-risk balance in the wider population. In the early post-authorisation phase, determining the added benefit of a medicine compared to standard of care, the remit of health technology assessment (HTA) bodies, and later relative effectiveness, could be a key application of real world data.

Privacy, security and transparency were overarching themes of the workshop. Frank and transparent exchange with patients and the public about how their data are used, how their privacy is protected and the potential benefits of data sharing must underpin projects to ensure success. The free flow of data, while protecting individuals' privacy, will be facilitated by the General Data Protection Regulation, coming into force in Europe in May 2018. This regulation focuses on the accountability of data controllers, making them responsible for compliance with data protection principles.

Workshop participants emphasised that patients and the public appreciate the value of sharing their data and that patients and patient groups play a pivotal role, not only in consenting to data use, but also in proactively driving data capture and providing informative patient perspectives.

EMA will continue to engage with stakeholders and will develop the skills and regulatory processes needed to ensure big data is harnessed to facilitate robust medicines assessments and complement clinical trial data. EMA is committed to an open and forward-looking vision, aligning the regulators with advances in technology for the protection of public health.

Videos and presentations from the workshop are available on EMA's website ■

Why a big data workshop?

"Big data enables us to generate a lot of conclusions; we have to be able to discriminate whether they represent causal relationships or spurious coincidence."

Professor Guido Rasi

Technological advances in both science and information technology are generating ever-increasing amounts of data on health and medicines. The objective of this workshop was to increase understanding of how big data will impact on our understanding of disease and facilitate medicines development, so that the regulatory community can identify opportunities and address challenges in its use for medicines decision-making. In his opening remarks, Professor Guido Rasi (Executive Director, EMA) emphasised the clear potential of big data to benefit patients. However, it is challenging to incorporate these data in a meaningful way into routine regulatory decision-making and importantly to understand how to determine whether the conclusions and associations arising from multiple analyses across varied data sets are causal and not simply spurious coincidence. Workshop participants included patient representatives, healthcare professionals, and representatives from government, industry, and academia, as well as regulators from across the globe ■

Defining big data

Workshop participants described the characteristics of big data as the '5 Vs': Volume, Variety, Veracity, Velocity and Value. Big data comprises massive data sets of far greater volume and variety than traditional data sets and may represent both breadth of data from large numbers of individuals and depth of data on each individual. Veracity refers to uncertainty of the quality and robustness of data from different sources. However, the sheer size and variety of big data may overcome a lack of quality in data sets. Today, data are accumulating at an unprecedented velocity and can be transmitted and analysed in real time. Big data provides value by enabling generation of information and knowledge to provide new insights, reveal hidden and rare associations, and increase efficiencies.

In medicines regulation and healthcare, the impact of big data is being felt in numerous ways. The availability of electronic health records from millions of patients provides greater understanding of the use, effectiveness and safety of medicines in clinical practice. Genomics data are redefining our knowledge of diseases enabling stratification according to genomic profile and personalised treatment. In addition, newly discovered biomarkers are facilitating earlier disease diagnosis and intervention, as well as tracking of treatment responses. In parallel, the widespread adoption of mobile devices and social media provide unprecedented access to patient-reported outcomes and lifestyle data and are already enabling the tracking of disease and spread of infections in real time. All these examples illustrate the potential of big data to benefit patients and the opportunities to incorporate patients' perspectives into medicines decision-making ■

Session 1—the big data landscape

Jean Georges from Alzheimer Europe opened the first session of the workshop, which aimed to provide an overview of the current state of the field. He noted the opportunities offered by big data, particularly in areas of unmet medical need such as Alzheimer's disease. Patients living with dementia appreciate that sharing their data is of vital importance to advance research and ultimately help other patients. In parallel, understanding and addressing ethical challenges such as informed consent and privacy is critical. The speakers in this session provided numerous examples of the use of big data in practice: Dr Lisa Latts (IBM Watson Health) spoke about use of machine learning algorithms that speed up discovery, monitor safety and create personalised and efficient healthcare; Nico Gaviola (Google Cloud Platform) gave the Google viewpoint on cloud computing and networks for big data analysis; Lauren Sager Weinstein (Transport for London) detailed the use of big data outside the healthcare sphere in London's transport network, and explained Transport for London's strategies for collecting and analysing big data and applying insights in order to benefit travellers.

During this session, speakers described the healthcare revolution that is currently underway due to the massive volume of healthcare data that has been generated (over 150 exabytes [1 exabyte=1018 bytes])[1] and its rapid rate of expansion (doubling every 24 months).[2]

1 Raghupathi, W. and Raghupathi, V. (2014) Health Inf. Sci. Syst. 2:3.
2 Marconi, K. and Lehmann, H. (Eds.) (2014) Big data and health analytics. CRC Press.

Emphasising this data volume, it was noted that there are over 100,000 clinical trials ongoing at any one time and 1.8 million new articles published in Medline annually.[3] It is consequently impossible for humans to be aware of all the knowledge and data being generated. In addition, there is a gap between what is known and what is implemented in clinical practice: it is estimated to take 17 years for an innovation to reach routine practice,[4] a time lag that might be shortened by big data applications. However, challenges to realising the opportunities offered by big data include the fact that 80% of generated data are unstructured and data are frequently noisy and are difficult to access due to storage across multiple silos.

"We want to combine the things humans are good at with the things computers are good at to improve the healthcare system."

Dr Lisa Latts

A fundamental concept is that data alone are not useful and value derives instead from the translation of data into knowledge that can be trusted as a basis for decision-making. Computing technologies that analyse big data are able to read data, analyse it and learn from human expert decisions in order to create insights. The underlying aim is to combine the things humans are good at—human-human interaction, creativity, compassion—with the things computers are good at—pattern identification, machine learning—to improve the healthcare system. To this end, Dr Latts described how IBM's Watson system has been used to create efficiencies in personal healthcare budgets and to monitor provision of these services for a local council, to identify medicinal safety signals and to speed up oncology discovery by identifying potential p53 kinases for further investigation.[5]

Big data also offers the possibility of moving from a culture of sickness to a culture of health, where data on the health and well-being of an individual are considered alongside clinical measurements to assess illness. Speakers highlighted that only 10% of the health of an individual is determined by routinely considered clinical factors, while 20% is determined by genetics, 30% by the environment and 40% by health behaviours, and that the availability of new data sets, for example from digital health apps, may allow consideration of all these aspects. Currently in medicines regulation, focus is principally on collecting data on disease whereas it would undoubtedly be valuable to also consider how a patient's wider health status and lifestyle choices influence their response to medicines.

However, the increasing volume of healthcare data presents significant challenges in terms of the computational capacity required to handle such large data sets. Nico Gaviola described how moving away from on-site IT silos to the cloud can provide capacity and security for storage and downstream processing of big data. Open source cloud storage and machine learning tools are preferable to proprietary models because the technologies that can gain trust and build a network of users will become the technologies of tomorrow. Open source tools for big data analysis should be put into the hands of as many people as possible including algorithms for users to query their own data and machine learning algorithms that users can train and test with their own data. Education of healthcare professionals and academics to increase awareness of data analysis tools is a key and necessary part of creation of user networks.

Strategies for data collection were discussed during the session. It was emphasised that data should not be collected just for the sake of it or because it may be interesting at some point but to respond to questions that will translate into benefit. It is therefore necessary to ascertain the knowledge that is missing by asking the right questions in order to focus resources on gathering the relevant data to answer these questions. To enable them to focus on the right questions, Transport for London phrases questions in the following form: As a [job title], I need [big data insights] so that I can [make

3 Talk about health blog (Sept. 2011) J. Clin. Oncol.
4 Slote Morris, Z. et al. (2011) J. R. Soc. Med. 104: 510.
5 http://time.com/3208716/ibm-watson-cancer/

a decision my job expects me to]. This helps to drive targeted data collection.

"Big data technologies are particularly powerful when they have a network of people adopting them, who can share how they've been using them and how they've been successful."

Nico Gaviola

The importance of forward planning for data collection was also stressed. In terms of the data needed, it is advisable not to focus only on things that can be measured now, but also to ask what important data cannot currently be measured and how they could be measured in the future. Forward planning also applies to operational systems. Data collection may not always be the main focus of operational systems, and is often instead a by-product of a system that benefits users (e.g. Oyster cards or contactless payment at Transport for London), but it should always be considered when systems are being established. In addition, when planning storage requirements for IT platforms, storage needed for data analysis as well as for raw data should be taken into account.

Throughout the session, data security was highlighted as a key priority. To maintain the trust of users, it is essential that organisations have systems in place to control access to data, keep it safe, and protect privacy ■

Key messages from session 1

- Vast amounts of healthcare data are continually being generated, offering huge opportunities but making it impossible for humans to keep pace with all the information.
- Harnessing of the potential of big data by researchers and regulators is hindered by the fact that it is often unstructured, noisy and inaccessible.
- Storing and processing disparate data sets is challenging and may be facilitated using cloud solutions and machine learning technologies.
- Deciding which data to collect starts by asking the right questions about the benefits sought and problems faced. Data should not be collected simply for the sake of it.
- Education of healthcare professionals and academics to increase awareness of data analysis tools will be key to maximise the impact of big data.

Session 2—big data meets medicines regulation: which data and when?

Professor Luca Pani from the Italian Medicines Agency opened this session focussed on defining how and when big data could support decision-making in medicines regulation. He highlighted that data only becomes useful when it is converted to information, which must in turn be translated to knowledge. He also emphasised the importance of thinking globally when working with big data.

The speakers in this session described examples of insights already gleaned from big data and the practical and strategic considerations needed to translate knowledge to clinical practice. Experiences to date in the field of pharmacogenomics were described by Professor Sir Munir Pirmohamed (University of Liverpool), who outlined learnings and challenges for regulators and healthcare professionals.

Dr Nicholas Tatonetti (Columbia University) spoke about identification of drug-drug interactions, an important aspect of regulation of medicines that may not be adequately addressed by clinical trials. He described use of data mining techniques to recognise drug–drug interactions and how to perform studies to produce reliable results that could inform medical practice.

At the level of healthcare systems, Dr Brian Kelly (Association of Clinical Research Organizations) and Nico Gaviola (Google) jointly presented on the transformations needed to enable precision medicine.

From a regulatory perspective, Professor Hans Hillege (EMA Committee for Medicinal Products for Human Use) described how exploitation of big data offers opportunities to support regulatory decision-making.

Professor Barend Mons (European Open Science Cloud, Leiden University) called for all data to be machine reusable to prevent data loss and exploit the potential of big data and also spoke about the open science approach of the European Open Science Cloud initiative.

Changes happening in the pharmaceutical industry in response to the new realities of digitisation and big data were detailed by Richard Bergström (European Federation of Pharmaceutical Industries and Associations, EFPIA). Closing the session, Roger Lim (European Commission DG SANTE) outlined European Commission initiatives in this field.

The field of pharmacogenomics has already significantly impacted medicines regulation and the amount of genomics data is continually increasing, coming from projects such as the 100,000 Genomes Project and Ubiquitous Pharmacogenomics. Pharmacogenomics information is already on the label of 15% of centrally authorised medicines, primarily in oncology, with 3.5% of medicines mandating a genetic test.[6] However, pharmacogenomics information on medicines labels may not always translate into clinical utility.[7]

6 Ehmann, F. et al. (2015) The Pharmacogenomics Journal 15: 201.
7 Wang, B. et al. (2014) JAMA Intern Med. 174: 1938.

Clearly, incorporating pharmacogenomics into medicines decision-making poses challenges to regulators, who must move towards making decisions on a medicine's benefit-risk balance based on data from small, stratified patient groups rather than at the population level. Incorporation of genomics may lead to regulators being asked to consider new types of trial design based on genetic mutations, such as umbrella trials (different mutations, one disease) and basket trials (different diseases, one mutation). Scenarios could arise where regulators may authorise expansion of a medicine's indication to add genotypes without the need for additional trials when an underlying common mechanism driving the disease pathology is known.

Incorporation of pharmacogenomics information into the medicine's label should ideally be done in a consistent way, ensuring patients benefit from available knowledge. Professor Pirmohamed described inconsistency in application of information, for example, testing for a HLA mutation (HLA-B*1502) that predisposes to carbamazepine hypersensitivity in Asian populations is mandatory, however, a predisposing HLA mutation in European populations (HLA-A*3101),[8] is included on the label for information only and no testing is required. Also despite a clear potential benefit to patients, genetic polymorphisms with the same effect size as renal impairment usually do not lead to advice or require testing in the medicines' label.

When medicines require genetic testing, issues arise with analytical validity of tests. It is necessary to ensure that tests accurately detect presence or absence of the variant being tested; for example, if tests from multiple providers can be used, they all must provide accurate results. In the future, as the field moves from testing for single genetic markers to testing panels of markers or next-generation sequencing, it will be important to ensure authorised tests are sufficiently reliable.

Ultimately in the clinic, the doctor's challenge is to consider all relevant information, including pharmacogenomics, and communicate and implement this information in a meaningful way during a typical 10 minute consultation time with the patient.

From pharmacogenomics, the workshop focus turned to identification of drug-drug interactions, another area where evidence from big data may in the future have significant impact. Drug-drug interactions can cause unexpected side effects and are responsible for a significant proportion of adverse medicine effects making their timely identification critical for medicines oversight. However, potential drug-drug interactions cannot all be systematically investigated in clinical trials and they are instead often detected following authorisation using observational studies or from spontaneously reported adverse drug reactions.

"When is enough enough? We have to start trusting that these algorithms are working, it's not possible to do prospective trials on every effect."

Dr Nicholas Tatonetti

It is acknowledged however, that observational data are messy and are confounded by multiple biases and noise from factors such as comorbidities and lifestyle choices, and analytical methods are needed to address this within adverse event reporting systems. Dr Tatonetti described a method called SCRUB (statistical correction of uncharacterised bias), which corrects for confounding factors in very large data sets.[9]

An additional issue is that the data may not include reports of a particular adverse effect, with the result that medicines associated with that adverse effect cannot be identified. To overcome this issue, machine learning is being used to identify severe adverse drug reactions from the combined observations of other measured effects in the same way that a disease can be identified by its signs and symptoms. For example, bradycardia, atrial fibrillation and

8 McCormack, M. et al. (2011) N. Engl. J. Med.364:1134.
9 Tatonetti, N.P. et al. (2012) Sci. Transl. Med. 4: 125ra31.

tachycardia could indicate the serious adverse effect of acquired long QT syndrome.

In the studies presented, drug-drug interactions were identified that cause diabetes-related adverse events (paroxetine, an anti-depressant medicine and pravastatin, a cholesterol-lowering medicine) or prolonged QT intervals (ceftriaxone, an antibiotic and lansoprazole for gastroesophageal reflux disease). The studies involved data mining of adverse event reporting systems to predict interactions, followed by corroboration with data from electronic health records. Laboratory models were then used to confirm the effect or provide evidence for a proposed mechanism of action.

During discussions, differing opinions on how to approach missing observations and selection bias were expressed. On one hand, since it cannot be known what observations are missing, it was proposed that work should focus on ensuring the data are rich before starting data analysis. An alternative approach is to perform analyses on the currently available data in order to generate hypotheses, followed by corroboration using another carefully selected data set and then further experimental validation. The feasibility of performing such analyses in the context of routine medicine regulation should also be considered.

The session moved to consideration of strategy with a presentation of the strategic developments necessary for healthcare systems to deliver precision medicine. To create a healthcare system in which treatments can be tailored to individual patients, access to and integration of a wide variety of data types such as clinical, laboratory, genomics, imaging, sensor and patient-reported data will be essential. However, such integration requires appropriate infrastructure and technology for collection and storage so that data from multiple systems can be extracted and aligned allowing the application of machine learning as well as traditional analytics. While bringing the data together in this way is challenging, once achieved analysis is easier and extremely fruitful both for research and patient care. However, reiterating the earlier point made by Nico Gaviola, it was pointed out that rapid advances in data accumulation and processing will necessitate the use of the cloud to store, access and enable the analysis of data because local infrastructure will be unable to keep pace with development. Examples of machine learning algorithms trained on large amounts of imaging data or genomics data were presented. These provided better results than doctors for identification of diabetic retinopathy from images of the eye (97-98%) or were superior to leading bioinformatics tools for analysing genomics data.

Data strategies of healthcare organisations will also be impacted by changes in health insurance reimbursement, which is becoming more merit or outcome based. In the US, for example, the MACRA Quality Payment Program ties Medicare payment rates to performance and by 2020 approximately 22% of a hospital's reimbursement will be tied to a clinical quality outcome. This is driving a need for hospitals to consistently collect high-quality clinical data to enable performance analysis.

Bringing the focus of the session back to regulatory science, a regulator's perspective on the opportunities for use of big data, with a particular focus on real world data, throughout the medicines lifecycle was presented. Opportunities are present before licensing (in development, scientific advice, generation of paediatric investigation plans and orphan medicine designation), during the evaluation process itself, and during the post-marketing phase (for pharmacovigilance and continuing evaluation of efficacy) where use rapidly increases and broadens. It was highlighted that relatively limited data are available at licensing (especially for orphan medicines which are often licensed with a requirement for post-authorisation studies and registries) and much of the data about a medicine become available in the post-licensing phase. Health technology assessment (HTA) bodies, with their focus on clinical added benefit and cost effectiveness also work in this post-marketing space where efficacy as shown in clinical trials does not necessarily translate into clinical benefit.

Discussions focused on the value of prospective registry and trial studies compared with retrospective analysis of big data. Increasing the inclusion of big data into medicines research is hampered by the heterogeneity in electronic

healthcare systems and policies across Europe. The results of observational studies are often not considered sufficiently robust but prospective studies may not be possible or may take a long time, as in the case of insulin glargine where observational, prospective studies after licensing took four years to establish that insulin glargine did not increase the risk of cancer.[10] The question was asked: how much work is needed before the knowledge from data analysis is considered satisfactory by regulators and the healthcare community, or when is enough enough? Eventually, the regulatory community will have to develop guidelines around acceptability and the extent of uncertainty that can be allowed, as performing prospective trials on every possible combination and effect is not feasible. Building trust, as much among scientists, regulators and clinicians potentially using big data as among patients and wider society is essential. Bad analysis of observational data must be avoided, whether this is big data or a trial dataset.

Returning to the topic of inaccessible, unusable data, workshop participants heard a call for all data to be made FAIR (Findable, Accessible, Interoperable and Reusable) to prevent the current massive loss of data and facilitate big data studies.

"80% of data is lost because it is not stored in a good repository."

Professor Barend Mons

FAIR is not a standard, but a set of principles to ensure data can be used, not just by humans but by computers. From January 2017, it is obligatory to use 5% of every Horizon 2020 project budget for good data stewardship. The data stewardship cycle starts with design and provenance and includes processing, preservation, infrastructure and analysis. Data provenance is important, especially for big data, in order to understand biases. With really big data, it was pointed out that patterns can be seen even if the data are not good quality, but knowing their provenance and context gives an understanding of how they might be biased.

Data stewardship and data analysis requirements will require training more people in these skills, as there are too few data scientists and bioinformaticians. Nevertheless, participants agreed that the majority of the challenges to achieving FAIR open data are cultural rather than technical.

Returning to the pharmaceutical industry perspective, participants were informed of changes the industry is undergoing to meet the disruptions that will come from the use of big data, increasing digitisation and the concept of new models of payment for packages of services or outcomes. Pharmaceutical companies are participating in the Innovative Medicines Initiative (IMI), the world's largest public-private partnership for health research, which runs several projects involving big data, including "Big data for better outcomes".

Pharmaceutical companies are also involved in the development of a system for serialisation of medicines in Europe in which, from 2019, each medicine pack will have a unique barcode. The barcode can be scanned by pharmacists to ensure the medicine is not counterfeit, but the system can also be used for reimbursement, pharmacovigilance and pharmacoepidemiology. Connecting the physical world of medicines with the digital one, such as through combining barcodes and wearable sensor technologies, has the potential to deliver big data sets of good quality.

Day one of the workshop closed with an outline of the importance of the digital single market, including health, in the European Commission's objectives. Commission big data initiatives include the eHealth action plan and Horizon 2020 funded research (4 projects: MIDAS, Evotion, BigO, PULSE) as well as public-private projects (Big Data Value Association and IMI). Future priorities on big data in public health, telemedicine and healthcare are available in a recent report.[11] It was noted that more detailed follow-up, particularly at Member State level, will be needed to ensure that these initiatives link to the priorities of HTA bodies and regulators ■

10 http://www.ema.europa.eu/docs/en_GB/document_library/Medicine_QA/2013/05/WC500143823.pdf
11 https://ec.europa.eu/digital-single-market/en/news/study-big-data-public-health-telemedicine-and-healthcare

Key messages from session 2

- Experiences with incorporation of pharmacogenomics information into medicines labels to date have revealed issues of clinical utility, test variability and inconsistencies in what information is presented in the label and when tests are recommended.

- Big data analysis can provide knowledge that may not be available from traditional studies such as clinical trials. But how can evidence be validated so that it is accepted by regulators and healthcare professionals?

- To perform precision medicine, organisations need to have lots of types of data stored in good infrastructures. Advancement towards such systems will be propelled by the growing movement of health insurers towards outcome-based reimbursement.

- Serialisation of medicines, where each medicine pack will have a unique barcode, will be implemented from 2019. This is a measure to prevent medicines falsification; however it offers possibilities for use in pharmacovigilance and pharmacoepidemiology although these are yet to be defined in depth.

- Data should be FAIR (Findable, Accessible, Interoperable and Reusable). From January 2017, 5% of every Horizon 2020 project budget must be used for data stewardship.

- Objectives and funding priorities of the European Commission favour the advancement of big data analysis for public health.

Session 3—how do we transform data into knowledge to support decision-making?

Session 3 explored how big data can be converted into knowledge of sufficient quality to support regulatory decision-making. Dr Thomas Senderovitz (Danish Medicines Agency), who opened the session, asked when regulators can trust knowledge derived from big data to inform robust decisions. Generating evidence from big data was described from the perspectives of the US regulator, the OHDSI consortium, and the pharmaceutical industry by speakers Dr David Martin (US Food and Drug Administration [FDA]), Dr Patrick Ryan (Observational Health Data Sciences and Informatics [OHDSI]) and Dr Bart Vannieuwenhuyse (European Federation of Pharmaceutical Industries and Associations). The potential of social media data and digital epidemiology to support the work of public health organisations was detailed by Professor Marcel Salathé (École Polytechnique Fédérale de Lausanne).

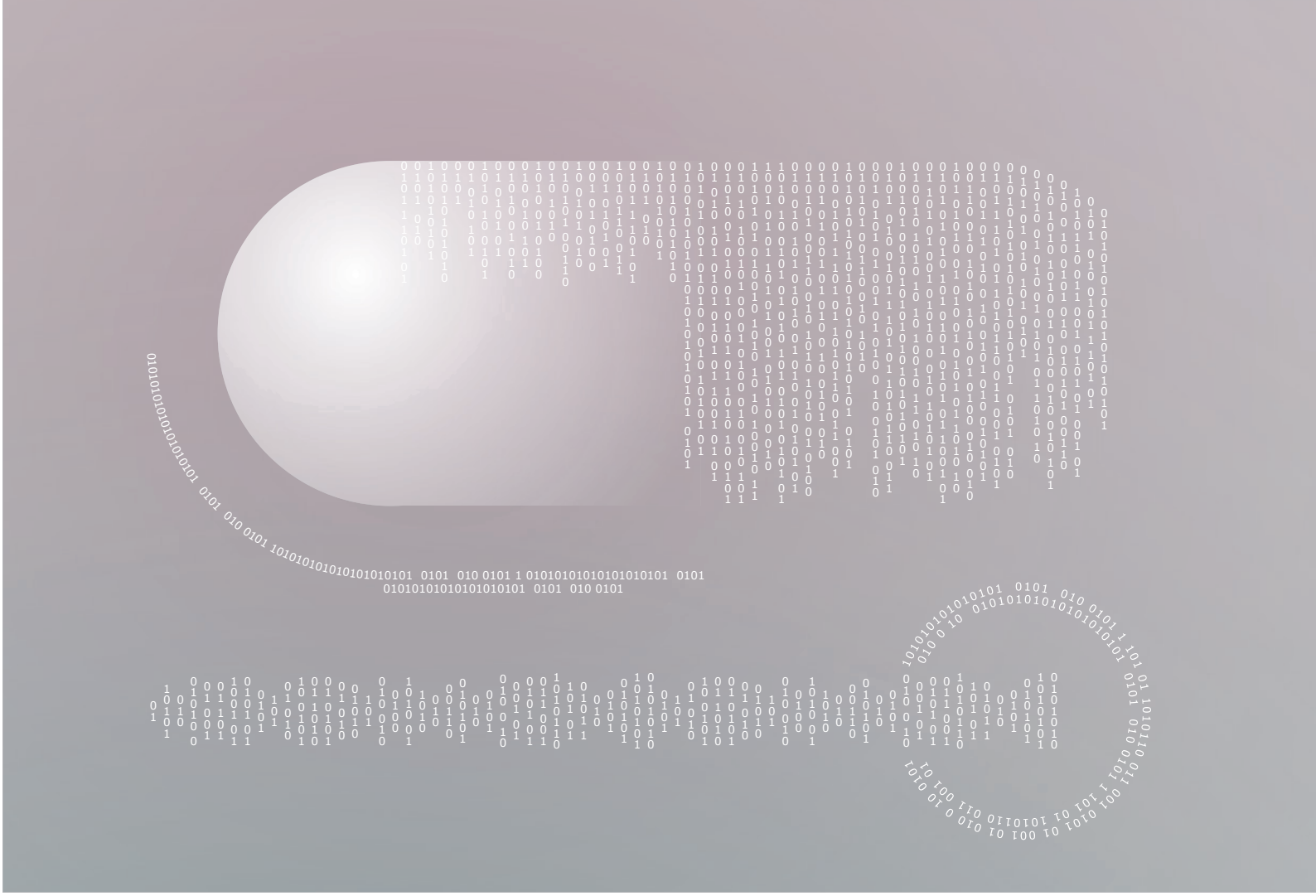
When assessing the attributes needed for evidence to support decision-making, FDA use a concept called 'sufficiency'. Sufficiency is represented by adequate data collected and analysed using appropriate methods and answering the question of interest to a satisfactory level of precision. FDA's ongoing work in the area of big data includes public consultations and pilot activities. FDA also intends to publish guidelines on use of real world data in regulatory submissions.

Currently, FDA has access to several big data sources, in particular from the Sentinel Initiative (FDA's electronic system for monitoring safety of medicines) and the Centers for Medicare and Medicaid Services (CMS, payer systems), which together comprise health data for over 200 million people. Sentinel provides centralised access to administrative data from 18 different health insurance sources across the US, delivering data on, for example, demographics, time under observation, hospital and doctor visits and pharmacy dispensing. These data can be complemented by data from electronic health records, which include additional clinical detail and lab/imaging results, but contain a lot of unstructured data and vary between institutions making the content hard to link. In addition, FDA is working on a mobile app with the aim of incorporating patient-reported data and covering areas where there is currently no visibility, such as use of over-the-counter medications. The app will initially be used to study medication safety during pregnancy.

To address the challenge of data integration, Sentinel uses a common data model that works across the multiple databases involved in the initiative. In this approach, data are converted into a common, structured data format that can then be used with modular programmes for various analyses. Use cases include describing medicine use or health outcome patterns, determining the incidence of health outcomes after exposures and inferential analyses comparing health outcomes in exposed and unexposed patients.

The Sentinel initiative promotes transparency by publishing studies, code and results to allow them to be reproduced or replicated on other data sets. The Regan-Udall Foundation, an independent organisation that supports the FDA, is based on public-private partnership and is working towards enabling users outside FDA to access data sources and methods.

Like Sentinel, OHDSI, an international, multi-stakeholder collaboration, covers multiple databases and uses a common data model to enable integration of data. The OHDSI network includes over 50 databases covering



over 660 million patients, and uses the OMOP (Observational Medical Outcomes Partnership) common data model, which converts data to a common structure, enabling its use for varied analyses including safety monitoring, comparative effectiveness research, clinical research and health economics. The data can be used to make observations for clinical characterisation (e.g. what happened to these patients?), or inferences including patient-level prediction (e.g. what will happen to me?) and population-level effect estimation (e.g. what are the causes?).

The power of the ODHSI model was highlighted in a recent study which aimed to provide a clinical characterisation of anti-depressant treatment of 250,000 patients.[12] Clinical guidelines for depression are very broad, and the study results clearly demonstrated that treatment is heterogeneous across geographies, health systems and time, and showed that there was no single preferred first-line treatment. Moreover, 11% of these patients had an individual treatment pathway that was unique to them alone.

Another ODHSI study underlined problems with the way evidence is currently produced from observational studies: from over 10,000 papers of observational studies in PubMed, 85% of exposure-outcome pairs are reported as statistically significant ($p<0.05$). In contrast, when OHDSI looked at 17 treatments and 22 outcomes for depression with 4 databases, just 11% of exposure-outcome pairs were significant ($p<0.05$). The difference may be due to observational study bias, publication bias or 'p-hacking' (collecting or selecting data until results show statistical significance). To address these issues, it was suggested that the observational research community should rethink how big data studies are carried out to ensure that they can generate reliable results that can be used as a basis for regulatory decisions. It is interesting to note that in the physics community, projects such as the construction of the Large Hadron Collider are

12 Hripcsak, G. et al. (2016) Proc Natl Acad Sci U S A. 113:7329.

much larger and more collaborative than projects in other scientific disciplines. Following the physicists' example, the observational research community should agree best practices for data analysis. Instead of picking individual study questions, the community could work together to objectively generate all the data needed for a particular research area.

"How do we come together to build the Large Hadron Collider of observational research?" **Dr Patrick Ryan**

For pharmaceutical companies, big data use occurs most often during discovery and manufacturing in the medicines lifecycle, whereas real world data may be used at the time of medicine authorisation and post-authorisation. During discovery, characterisation of compounds and their known target affinities can predict affinities for new targets in silico and be used to suggest compounds for screening. This was described as being analogous to internet sellers using algorithms to suggest purchases to customers based on their preferences. In manufacturing and operations, big data plays a role in monitoring production processes and integrating supply chains.

Harnessing big data for discovery is the aim of European initiatives such as the IMI project, the European Medical Information Framework (EMIF). EMIF is developing a European framework for sharing of health data, and data sets are combined utilising the same OMOP common data model as ODHSI. EMIF covers data from 40 million patients and initial projects are focussing on prediction of biomarkers of early onset and progression in Alzheimer's disease and biomarkers of complications in metabolic disease.

In Europe, a network of hospitals which has grown from the IMI project EHR4CR (which ended in 2016) is being developed in collaboration with several pharmaceutical companies. The network will apply a common data model to locally stored data, enabling remote research queries to be sent to participating hospitals across Europe. This data analysis will be used to improve costly problems faced by clinical trials in recruiting patients on time by better informing trial feasibility, trial design, eligibility criteria and site selection. The collaboration even aims to avoid the need to enter data twice, often encountered during clinical trials, by integrating existing data from hospitals participating in trials.

The session also highlighted the potential power of social media data and digital epidemiology in healthcare. A lack of opportunity for healthcare professionals and agencies to gather patient information, due to factors such as lack of healthcare infrastructures in some parts of the world, short 10 minute consultation times, or data not being accessible to authorities, create a 'bandwidth problem' in traditional epidemiology. In contrast, digital epidemiology takes advantage of new data streams of text, image, video, sound, location and biological data, shared constantly via mobile apps, though the data may be noisy. The technology will soon be available to all: currently in the US, more time is spent on mobile apps than any other methods of accessing the internet and availability of mobile broadband will approach saturation in next decade.

"Traditional epidemiology has a bandwidth problem."

Professor Marcel Salathé

Examples of the application of digital epidemiology include use of mobile data to see how travel affects malaria spread, influenza forecasting using publicly available Wikipedia access logs and use of Twitter data for pharmacovigilance and to monitor vaccine uptake. A study of Twitter sentiment during a H1N1 vaccination programme correlated with vaccine coverage reported later by the Centre for Disease Control.[13] Such information could be used to guide decision-making in real time.

It is clear that digital epidemiology must become an integral part of public health institutions. The biggest challenge is data acquisition as health

13 Salathé, M. and Khandelwal, S. (2011) PLoS Computational Biology, 7: e1002199.

institutions are increasingly out of the loop, and no company alone has all the relevant data. A 'WHO of data' should be established to take data out of silos so that it is possible to see the 'big picture'.

At the end of this session, reduction of bias was discussed, and it was suggested that performing randomised trials within big data can overcome bias. This is an approach that is ongoing with a trial using Sentinel
called 'IMPACT Afib', which aims to improve treatment with oral anticoagulants in patients with atrial fibrillation. It was acknowledged however that in Europe there are many barriers to doing randomised trials within electronic health databases due to varied legislation, systems and access across borders. Such trials also need appropriate statistical paradigms, since statistical methods for small datasets are not appropriate for big data ■

Key messages from session 3

- Many networks are using a common data model to enable use of multiple databases across organisations. Data control stays at the local site but data can be accessed as needed in a common format for analysis to respond to questions posed centrally.
- To increase reliability and realise the potential of observational evidence, best practices and open-source tools for analytics should be established.
- Collaboration across networks in the observational research community could enable generation of knowledge for entire research areas to meaningfully inform medicines decision-making.
- Big data is used throughout the medicines life cycle, especially facilitating discovery. Real world data is increasingly important at authorisation and during the post-authorisation phase.
- Digitalisation and social media are now ubiquitous and data can be analysed using machine learning. If such digital epidemiology data are acquired and integrated by public health organisations, they have the potential to be used to support public health decision-making in real time.

Session 4—reconciling big data and privacy: legal safeguards for unleashing technological innovation

Big data brings big opportunities but also big responsibilities. Privacy and security measures were mentioned by speakers throughout the workshop, and these were the topics in the spotlight for the final session opened by Professor Jon Snædal (World Medical Association [WMA]). WMA has recently adopted the Declaration of Taipei, which lays down ethical considerations regarding health databases and biobanks. In this session, Sophie Louveaux (European Data Protection Supervisor) presented the General Data Protection Regulation, outlining the opportunity it offers to build trust with patients. Sharing knowledge gained from involvement in hundreds of registries and patient studies, Professor Ronald Brand (University of Leiden) spoke about privacy principles and security and data encryption methods for protecting personal data. The experience of the UK biobank, including ethical issues and data governance considerations, was shared by Baroness Helene Hayman (House of Lords, UK Biobank Ethics and Governance Council). To close the session, the value of patient perspectives was highlighted by Julian Isla (Dravet Syndrome Foundation), who presented Wacean, a patient-driven data capture tool.

The new General Data Protection Regulation (GDPR) comes into force in Europe in May 2018. GDPR aims to promote a sustainable approach to data protection in the digital era, facilitating free flow of data and the internal market, while protecting individuals' privacy. The regulation defines health data as personal data related to the physical or mental health of a natural person, which can include lifestyle data if they are used to determine health conditions.

GDPR incorporates rules on consent, which must be free, informed and specific, although this can pose problems with big data analyses where it may not be known at an early stage what patterns will emerge and hence the exact purpose for which the collected data will be used. However, there are specific legal alternatives to consent that may be used when necessary for scientific research. Principles of data minimisation and privacy by design and by default are also incorporated in the regulation, as well as the principle of accountability of data controllers. Accountability comes with collection and processing of data, and the regulation requires measures such as documentation, performance of a data protection impact assessment, security requirements and designation of a data protection officer. Accountability brings public trust and readiness for the future and the regulation should be seen as a tool that helps to improve the design of studies and the quality of data.

Professor Brand described the privacy principles and security measures implemented at the Leiden University Medical Centre, where multiple registries and studies of patient data are designed, maintained and analysed. It is recognised that there is a potential conflict between the needs of data collection to support patient care and those appropriate to research. Studying health and illness requires the researcher to follow the patient through time and space. This will inherently invade the patient's privacy but a balance can be established by following principles of necessity (only collect the data needed for the questions that will be asked), proportionality (only collect as much data as needed) and subsidiarity (use less sensitive data where possible). Data registries should be designed to require

informed consent (this can be via informed opt-out rather than informed opt-in), appropriate security (access limitation, intruder detection, encryption of identities separately from database access rights), certification, and transparency.

"We have to be frank with the public, and show them how we are working, to maintain their trust and support." Baroness Helene Hayman

The UK Biobank is also an example of a health resource that provides a blueprint for addressing the challenges of collecting and maintaining personal data. At UK Biobank, over 0.5 million people have provided samples and consented to access to their health data. Consent can be withdrawn at any time, however this has not happened to any significant extent as trust has been maintained with participants. The latest stage of the project involves MRI imaging on 100,000 participants, making it the largest imaging project worldwide.

The underlying ethical principles of a project such as UK Biobank or other big data projects are not new, however it is necessary to learn how to apply them in the big data context. It is critical to take appropriate measures from the outset, otherwise public confidence can be destroyed to the detriment of the project.

In its work, the UK Biobank encounters many ethical issues, particularly regarding when to give health information to participants on incidental findings and how to protect participants' best interests, and it has a protocol for how to respond to ethical questions. It was noted that public opinion can vary over time, and there may be a periodic need to revisit decisions on patient protection and the principles behind them. However, the key to maintaining trust is always to be frank and transparent with participants and the public about what is being done with their data.

The session and the workshop presentations were brought to a close with a reminder of the valuable contribution of individual patients and patient groups. EMA has led the way in patient

involvement by including patient representatives in its scientific committees including the orphan medicines committee, COMP, responsible for medicines for rare diseases. However, for rare diseases such as Dravet syndrome (a seizure disorder) there is a lack of accessible data for regulators and doctors to use to make decisions. The development of the Wacean data platform has allowed patients and carers to collaborate in collecting and storing information on their medicines use and condition in an accessible, portable format. Personal information is carefully separated from clinical information, but the system maintains the centrality of the patient, allowing the individual to compare their outcomes to those of the aggregate in real time. The tool is now being used to support development of new treatments, in collaboration with the regulator. In the context of rare diseases, affecting relatively few patients, big data is characterised by depth of information on individuals, rather than information on large numbers of individuals. However, as we identify more and more genetic subsets within commoner diseases, big data models developed for rare diseases may spread into the mainstream ■

Key messages from session 4

- The GDPR comes into force in 2018 and aims to facilitate free flow of data while protecting privacy. The regulation covers consent, data collection principles and security requirements and introduces the requirement of accountability for data controllers.
- Principles of necessity (only collect the data needed for the questions that will be asked), proportionality (only collect as much data as needed) and subsidiarity (use less sensitive data where possible) should be followed for data collection and use.
- Privacy, security, ethics and transparency measures should be implemented from the outset of projects to maintain trust of participants.
- Patients can provide valuable data and insights to regulators and can drive innovation and data capture.

Panel discussion — brainstorming big data

"This is not a European challenge, it is a global challenge, and the more we work together and share experiences across borders, the better."

Dr Thomas Senderovitz

Dr Pierre Meulien (IMI) chaired the panel discussions. He remarked that IMI must join the dots between their big data projects ensuring that they do not add to the fragmentation in the field, and that future projects should build on existing achievements. IMI will continually strive to include stakeholders including EMA, HTA bodies and patients. Big data projects should be international, as building data sets should be done in a global forum. Dr Meulien introduced three questions for the panel discussions, on data integration, trust in data, and managing the pace of change. The discussions around these topics are summarised below.

How can we successfully integrate data?

Panel: Professor Miriam Sturkenboom (Erasmus University Medical Centre), Professor Sir Munir Pirmohamed (University of Liverpool), Dr Andrew Leach (European Bioinformatics Institute)

The panel considered that the barriers in the area of data integration can be overcome. However, the principal challenge is to identify the questions the data should be used to address and to have the imagination to understand what can be achieved with the data. This requires looking outside silos to understand what data are being generated in other fields and how they can be used to maximum potential. For example, information on patient experiences is available in many locations, and should be integrated to create health records containing complete patient journeys. Trust between different disciplines is needed, so that when data are integrated from other areas, its value is appreciated. It is also necessary to continue to develop better methods of analysis and data modelling.

The panel discussed the issue of discriminating between correlation and causality, and agreed that increasing understanding of biological mechanisms will help to address these issues allowing causality to be more reliably inferred.

Appropriate and considered communication is fundamentally important to gain public trust. It is necessary to communicate better, in a patient-specific format, what big data means in healthcare, including the types of data, what it is used for, the conditions under which it can be used and the potential benefits that may result.

The challenge of how to measure impact was briefly discussed during this panel discussion. It was suggested that as big data is increasingly used to develop new medicines and disease stratifications, ways to determine the impact of measures on public health should be considered in parallel.

How can we build trust in the data?

Panel: Rob Hemmings (Medicines Healthcare and Product Regulatory Agency), Dr Olaf Klungel (University of Utrecht), Professor Stephen Evans (London School of Hygiene and Tropical Medicine)

Trust in the evidence coming from big data will depend on where in the medicines life cycle the evidence is used. Often, especially if looking early in the lifecycle, data don't always have to be of perfect quality because fluctuations are smoothed out by having large amounts of data.

During the clinical phase, it may be possible, depending on the context, to complement a clinical trial with observational data. In the post-marketing phase, where safety is often the objective of studies, establishing causality is more difficult. In addition, treatment guidelines often make post-marketing observational studies even more difficult because treatment is not only not allocated randomly, but if treatment is determined by the guidelines, without external data or very strong assumptions, effects of patient characteristics and treatments cannot be distinguished.

Collection of data from clinical trials follows quality standards including Good Clinical Practice standards provided by the International Council on Harmonisation and CDISC (Clinical Data Interchange Standards Consortium) standards. Regulators are thus confident of provenance of clinical trial data and the control of its collection, which builds trust. From a regulatory perspective, EMA has some experience of other data sources being presented by companies to contextualise trial results or justify trial design or even replace a trial, especially in the context of unmet clinical need. In line with this, the recent EMA adaptive pathways pilot aimed to facilitate development of medicines that challenge the conventional model. However, it has proven difficult to validate that other data sources can provide information as robust as traditional clinical trial data. While it is often easy to agree with companies on what are the questions to be answered, it is more difficult to agree the data source to use and its quality, provenance and reliability, before even starting to consider further details of design and analysis. As regulators, EMA will document best practice and offer regulatory guidance, as well as possibly guidelines on standards for reporting. EMA already has several mechanisms in place for dialogue, such as the Innovation Task Force, Scientific Advice and qualification procedure.

How can we manage the pace of change?

Panel: Dr Thomas Senderovitz (Danish Medicines Agency), Dr Hugo Hurts (Medicines Evaluation Board), Dr Bart Vannieuwenhuyse (European Federation of Pharmaceutical Industries and Associations), Dr Lisa Latts (IBM Watson Health)

It was universally agreed that the regulatory network is facing increasing challenges and it cannot afford to ignore the opportunities offered by the exploitation of big data. The first point made by the panel was that flexibility will be key to managing change, although this is not a characteristic traditionally associated with regulation.

For regulation, randomised clinical trial data will continue to be used, with the addition of different types of data as the field progresses. Integration of big data evidence will enable regulators to see the bigger picture, rather than just a narrow dossier, and this will drive better decisions. It is sensible to focus on implementation rather than innovation, creating the right infrastructures and building methodological rigor around big data, as has been done for data from randomised clinical trials.

A significant bottleneck is the lack of people with the appropriate skills, specifically those qualified in data science with life science or medical backgrounds. This is particularly true within regulatory agencies, and the point was raised that the FDA has in-house data analysis skills, a resource which could also be considered in Europe.

Panellists agreed on the value of an open science approach. Public-private partnerships will be important for implementation, but an open science approach rather than commercial models is necessary at this time to advance the field to maturity and stability. Data should be accessible to both companies and researchers ■

Next steps from EMA's perspective

"We can, and we must, seize the opportunities for public health coming from big data" **Professor Guido Rasi**

Professor Rasi, EMA's Executive Director, thanked the workshop organisers and participants for the fruitful and thought-provoking talks and discussions. To conclude the workshop, he offered his thoughts on the impact of big data on decision-making for medicines and next steps for the regulatory world.

- The tools and data described in this workshop will not replace randomised clinical trials, but will improve clinical trials and also complement trial data, supporting decision-making on medicines. Regulators need to better understand how to use these types of information to support future decision-making and must be able to differentiate between causality and coincidence.
- The regulatory tools should be put in place to enable use of big data analytics in the regulatory context, so that we are ready to take full advantage of the value of this type of evidence.
- Data access and use should be for the common good and should not be commercialised.
- The regulatory network should reflect on current investments in telematics and IT infrastructures to ensure that resources are focused on the right areas for the future.
- The knowledge generated from large data sets and new computing technologies is as relevant for the work of HTA bodies as it is for regulators. The boundary between regulators and HTA bodies may be naturally dissolved by the coming wave of evidence.
- EMA will continue to engage with its regulatory partners and with stakeholders to promote an open and forward-looking vision, aligning the regulators with advances in technology ■

European Medicines Agency

30 Churchill Place
Canary Wharf
London E14 5EU
United Kingdom

Telephone +44 (0)20 3660 6000
Facsimile +44 (0)20 3660 5555
Send a question www.ema.europa.eu/contact

www.ema.europa.eu

Identifying opportunities for 'big data' in medicines development and regulatory science
EMA/740359/2016
Stakeholders and Communication Division