Much has been written about the potential of data mining – big data – to transform drug development, reduce uncertainty, facilitate more targeted drug discovery and make more personalised medicines and accelerated access to medicines a reality. The collection of data has always been central to medical science. What is new is the scale of the data and the evolving technology used to evaluate it. Here we consider what is driving this new data revolution in life sciences and healthcare.

**A Changing Industry**

The business climate for BioPharma companies has changed dramatically in recent years. Globally, healthcare systems are changing as governments implement policies to ensure healthcare costs are sustainable. Patent expiries, declining research and development productivity, pipelines of more high-risk molecules with often smaller potential markets and changing incentives (including moves to more outcomes-based pricing agreements) are pushing companies to review their drug development strategies and adopt more sustainable business models.

In parallel with this, a high rate of attrition continues, as drug developers are often unable to predict accurately the efficacy and safety of new medicines early in the development process, and the number of late stage failures continues to drive up the cost of developing new treatments. Increasing regulatory demands and the need to demonstrate effectiveness over existing treatments to support reimbursement are adding further complexity and cost to the development process.

**What Is Big Data?**

The term ‘big data’ refers to the use of technology to enable the capture, storage, distribution, management and analysis of a titanic amount of diverse data in real time. Big data is big in terms of the quantity and variety of data that are available to be processed and the scale of analysis (‘analytics’) that can be applied to it, ultimately to make inferences and draw conclusions. The data itself does not possess inherent value in the absence of a means to make sense of it.

**What Does Big Data Offer the Life Sciences and Healthcare Sector?**

Within the life sciences and healthcare sector, data already forms the basis of many decisions; clinical data defines whether a drug is effective, and advanced data models are used for cost-effectiveness decisions made by health-technology assessment (HTA) bodies.
In the current business climate there is a growing drive to capitalise on the increasing patient and health system data available and the ever more sophisticated data analytics techniques to revolutionise healthcare innovation, as the industry looks for ways to accelerate drug development and improve its competitiveness. These tools are seen as having the potential to increase productivity at all stages of the drug discovery, development and delivery process. Such data-driven approaches are also seen as a possible way to improve efficiency by enabling drug candidates to fail faster.

Data from a wide variety of sources needs to be combined for processing to reap the potential rewards. New insights may be found by mining gene expression data, clinical data, and data from third-party sources such as electronic patient health records, medical images and biobanks. Medical equipment and technology, monitoring everything from heart rate to blood chemistry, are now being networked and connected to electronic patient records, personal health records and other healthcare systems with the potential for real-time analysis of these data streams. In addition, advances in communication technologies such as mobile medical applications are supporting active patient engagement in their own healthcare, allowing patients to communicate in real time about their experiences with particular drugs and treatment regimes, facilitating the shift to more adaptive approaches to drug approval and reimbursement.1

One project worth mentioning is EUResist,2 in which data scientists use analytics tools to segment patient populations at an increasingly granular level. Such an approach has helped scientists to make huge advances in individualising treatment for HIV, where different combinations of drugs have varying effectiveness across patients.

Also in the United States, Ayasdi3 (which we are told means ‘to seek’ in Cherokee) is using advanced data analytics, amongst other things, to find cures for cancer. The company is working on complex algorithms to unlock value in the human genome. It has used these algorithms on breast cancer data collected over a 15-year period to identify a sub-group of patients who have a higher chance of survival based on their genetic profile.

Challenges and Risks: Striking a Balance between Privacy, Security and Innovation

These new capabilities to gather, analyse, disseminate and preserve vast quantities of data raise new concerns about the nature of privacy and the means by which individual privacy might be compromises or protected. In Europe, the data protection rules that are currently being formulated mean that the issues relating to big data, health research and privacy are under discussion. Two basic principles in the current text are the ‘right to be forgotten’ (when a data subject no longer wants its data to be processed and there are no legitimate grounds for retaining it, the data must be deleted) and ‘informed consent’. Both, in different ways, fetter the ability of big data initiatives to ‘profile’ individuals.

2) EUREsist prediction system: http://engine.euresist.org/.
3) http://www.ayasdi.com/.
From an ethical perspective, it is essential that individual patients (data subjects) are adequately informed of the current and future use that might be made of their data and also the advantages that they may derive by making their data available for biomedical exploitation. The same data and analytics that provide benefits to individuals and society, if used appropriately, can also do potential harm. For example, large-scale analysis of research on a disease, together with health data from electronic medical records and genomic information, might lead to better and timelier treatment for individuals but also, possibly, to inappropriate disqualification from insurance or jobs. Yet how can a data subject grant consent for as yet unknown purposes? Can it ever be known what information may later be extracted from any particular collection of big data? Probably not, because that information may result only from the combination of seemingly unrelated data sets and the algorithm for revealing this new information, both of which may not even have been available at the time the data was collected. Although the purpose limitation requirement serves to protect data subjects’ rights, if a company has to obtain informed specific consent for each new application of the data this could severely hamper the unlocking of the potential benefits that big data has to offer life sciences and healthcare.

The right of data subjects ‘to be forgotten’ seems to imply some sort of ownership of data relating to them, and, therefore, the right eventually to donate or sell that data. There are currently instances where others use that data, usually without the patients’ knowledge or consent, to earn profits. For example, there is already a growing market for data on individual prescriptions. Patients currently do not participate in the revenues generated from their respective data being added to large repositories. Regulation aimed at letting patients participate in some way in the financial advantages derivable from their data is a complex topic, and beyond the scope of this editorial, but there is a need for public debate on this issue.

It is vital that organisations employing data analytics understand who owns the data that they are using and whether they have a right, or licence, to use it. Data used in big data projects may be protected by copyright and/or database rights, provided there has been an intellectual effort in creating a copyright work out of data and/or a substantial investment has been made in obtaining, verifying and presenting the data. A distinction needs to be made between the data and the products and services developed through analysis of the data, as no direct IP rights are derived from the raw data as such. IP rights attach to the product or service deriving from the analytic work performed on the data; however, it is not clear how the ‘right to be forgotten’ may impact on these products and services. What if some individuals, whose data is part of the underlying data set used to produce a product based on patient data (for example, an in silico disease model that has been developed by analysing, and in a novel way aggregating a data set), request the deletion of their data: must the effects of their data on the product be removed and does the data set have to be reanalysed? If the product has been sold on, or if the data is totally anonymised, this may no longer be possible.
The current and the proposed new European data protection legislation also highlights the importance and desirability of anonymisation of data, but recognises that research value of the data may be diminished if the data is fully anonymised. In healthcare, veracity of data is essential; the more information that could potentially be used to identify patients is removed, the lower the data's veracity and thus its clinical value. Perhaps a patient's name may not be clinically significant, but age, gender and blood type could potentially be used to assist in identifying patients, while also having obvious clinical relevance.

However, it is important that it is not just the nature of the data that influences the level of protection afforded, but also the intended use of that data, and the potential risks implied by its usage. EU citizens have shown time and time again in surveys that their concerns over data security relate not to the use of their data per se, but to who will use it, and who might have access to it in the future. A harms-based approach has been suggested, focusing more on the regulation of possible misuse of data, rather than on limitations of usage, and may be the most appropriate approach for striking a reasonable balance between privacy, security and innovation.4

According to a recent paper, ‘Big Data and Privacy: A Technological Perspective’,5 prepared for the US President by the President’s Council of Advisors on Science and Technology (PCAST), some techniques for privacy protection used as supplementary ways to reduce privacy risk now do not seem sufficiently robust to be a dependable basis for privacy protection where big data is concerned. For a variety of reasons, PCAST judges anonymisation and data deletion to be in this category. Anonymisation may remain somewhat useful as an added safeguard in some situations but approaches that deem it, by itself, a sufficient safeguard need updating. The report also recommends that the framework of notice and consent is becoming unworkable for some of the reasons discussed above.

The other challenges of big data lie in capturing and storing the information securely, devising the tools to analyse and manage it, including standardisation and integration, and the challenge of open access, which is making information not just available but also readable and usable.

The question of interoperability of medical data systems is widely discussed as a challenge, and there are already attempts to deal with it such as the European Union's epSOS (European Patients Smart Open Services) system, which transmits medical data from country to country, and translates it from one language to another. There is also Scotland's Emergency Care Summary database containing a summary of demographic, allergy and medication information for 5.5 million people. It allows healthcare professionals access to important patient information in emergency and unscheduled care situations.

Extracting value out of healthcare data for all its beneficiaries (clinicians, clinical researchers, pharmaceutical companies, healthcare policy-makers and so on) will require new ways of working and, most importantly, new levels of transparency and openness. The Innovative Medicines Initiative (IMI) is an example of an EU-funded programme that aims to improve drug development and regulation through the use of pooled data. Unprecedented levels of data-sharing between public and private partners under IMI have resulted in the launch of a new in silico toxicology prediction software (eTOX) that could significantly improve the ability to predict the safety of new medicines and could play a direct role in reducing attrition rates.

Developing global transparency policies are also intended to facilitate wider access to clinical data, making data mining possible. Many companies that have traditionally been intensely protective of their intellectual property are devising policies to share drug trial data and other data. In addition, the new Clinical Trials Regulation and proposed new European Medicines Agency (EMA) policies will ensure that clinical trial data is more publicly available. Within these policies it is recognised that there is a need to balance the benefits of openness with legal and commercial requirements to protect patient and commercial confidentiality and safety. In the United Kingdom, Patients4Data supports openness and is campaigning about the power of patient data to save lives. Open source models for research are also becoming more prevalent. Recently, GlaxoSmithKline and Novartis have deposited data on compounds active against the malaria parasite with the European Bioinformatics Institute. How this open source model will fit within the traditional protected IP business model remains to be seen.

What Lies Ahead?

How the complex issues discussed above are reflected in the ongoing big data debate is still to be ascertained. The power inherent in big data is that it has the potential to provide the means for decision support across all aspects of healthcare, ranging from assessing the safety and efficacy of drugs to carrying out health technology assessments. Prevention, diagnosis and treatment could be refocused from the population-level, one-size-fits-all paradigm, to a personalised approach: ‘the right drug, for the right patient, at the right time’.

It is clear that, from the patient perspective, establishing trust is essential if the techniques and tools of big data are to be successfully applied to health. The technology is at hand to apply big data to health, but there must be a public debate about the risks associated with data-sharing. In the United Kingdom, the Nuffield Trust on Bioethics is currently exploring the ethical issues relating to developments that facilitate the collection, linking, use and exploitation of data relating to individual people. Concerns about big data reflect both the substantial increases in the amount of data being collected and associated changes, both actual and potential, in how they are used. However, it is important that data protection laws are not too prescriptive, since the key to extracting value from big data is to be able to apply new tools and to conceive new methods and approaches for converting data into insightful information.
As stated in the recent OECD Report on ‘Supporting Investment in Knowledge Capital, Growth and Innovation’, in order to reap the potential value from big data, not only will clinicians and researchers have to acquire big data analytics skills and services, but a framework needs to be developed for data repositories which adheres to international standards for the preservation of data, sets common storage protocols, protects the integrity of data and metadata, establishes rules for different levels of access, and defines common rules that facilitate the combining of data sets and improves interoperability. This framework could, some day, render some of today’s data protection rules and procedures unnecessary.