Received: 16 August 2015 | Revised: 27 July 2016 | Accepted: 21 September 2016

DOI: 10.1002/gepi.22012

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Use of big data for drug development and for public and personal health and care

Lada Leyens[1]* | Matthias Reumann[1,2]* | Nuria Malats[3] | Angela Brand[1,4]

[1]Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT), Maastricht University, Maastricht, the Netherlands

[2]IBM Research – Zurich Laboratory, Rüschlikon, Switzerland

[3]Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

[4]Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands

Correspondence
Lada Leyens, Hallerstrasse 7, P.O. Box, 3000 Bern 9, Switzerland.
Email: lada.leyens@maastrichtuniversity.nl

*These authors have contributed equally to this work.

ABSTRACT

The use of data analytics across the entire healthcare value chain, from drug discovery and development through epidemiology to informed clinical decision for patients or policy making for public health, has seen an explosion in the recent years. The increase in quantity and variety of data available together with the improvement of storing capabilities and analytical tools offer numerous possibilities to all stakeholders (manufacturers, regulators, payers, healthcare providers, decision makers, researchers) but most importantly, it has the potential to improve general health outcomes if we learn how to exploit it in the right way. This article looks at the different sources of data and the importance of unstructured data. It goes on to summarize current and potential future uses in drug discovery, development, and monitoring as well as in public and personal healthcare; including examples of good practice and recent developments. Finally, we discuss the main practical and ethical challenges to unravel the full potential of big data in healthcare and conclude that all stakeholders need to work together towards the common goal of making sense of the available data for the common good.

KEYWORDS

bioinformatics, healthcare, health systems, personalized medicine, data commons, public health, public health genomics, structured data, unstructured data, drug development, safety monitoring, health policy, comparative effectiveness research, knowledge visualization, cognitive computing

1 | INTRODUCTION

"Without the right analytical methods, more data just gives a more precise estimate of the wrong thing" (Ellenberg 2015).

The use of data in public health applications and in drug development and monitoring has been part of standard practice for a very long time. Epidemiologists have studied the patterns, causes and effects of health and disease in populations for decades combining different type of data to do so (Chiolero, 2013); biologists have evaluated data from complex biological systems to develop predictive models of disease for decades (Boyle, 2013); biostatisticians have analyzed clinical trial data to confirm or reject the studied hypothesis since the 1950s (the first randomized controlled trial of streptomycin was developed in 1946) (Bhatt, 2010); etc. So the following questions arise: "What is different now? Why is a there a sudden interest and hype in the use of 'big data' for healthcare?"

The new potential does not lay in the use of data per se but on the increased amount of information we are gathering in all fields and more importantly on the increased accessibility and "exploitability." New tools are being created to store, structure, analyze, and exploit these data; and the advances in technological capabilities enable more effective and smarter analysis of larger quantities of data simultaneously. All these aspects combined allow for the analysis and combination of new types of information (e.g., social media) that previously could not have been imagined to play a role in public health, drug development, and monitoring (Broniatowski, Paul, & Dredze, 2014).

This increased wealth of data is coming from many different sources and developments in areas that are not linked to each other, and not necessarily linked to health (see Table 1 for a snapshot). Some examples are the wider implementation of electronic health records (EHRs) in hospitals and healthcare providers; the faster and cheaper DNA sequencing methods that are delivering large amounts of genetic information daily; the vast information shared by patients and citizens through social media and the web; readings from remote sensors and devices measuring vital signs; data from insurance claims and billing records (depending on each regional/national

TABLE 1 Potential Type of Data in EU Healthcare [Modified From (Szezák et al., 2014) for the EU Setting]

Type		Description	Main source	Structured/ unstructured
Clinical	EHRs	Patient level clinical data	Hospitals and HCPs	Structured, unstructured
	Registries	Basic medical information specific to one disease/therapeutic area	Patient/health associations	Structured
	Diagnostics and biomarkers	Results from diagnostic tests and biological indicators of disease status or treatment	EHRs	Structured
Healthcare utilization	Insurance claims	Medical information and health utilization information from claims (depending on national healthcare system)	Public or private insurance providers	Structured, unstructured
	Admission, discharges	Summary of administrative information	Hospitals and HCPs	Structured
	Drug orders and sales	Records of drugs sales and revenue	Distributors, pharma companies, HCPs, pharmacies	Structured
	Clinical and pharmacy dispensing	Dispensing data from medications administered in hospitals and pharmacies	HCPs, pharmacies	Structured, semistructured
Biological	"-omic" information	Not related to clinical diagnosis (e.g., from biomedical research projects)	Research institutions, international molecular databases	Structured, unstructured
	Other biological information	Data from biomedical research	Research institutions, international molecular databases	Structured, unstructured
Drug development	Clinical research	Clinical trial design parameters and results	Pharmaceutical companies, regulators, international clinical trial repositories, biomedical journals	Structured, unstructured
	Safety and Pharmacovigilance	Adverse drug reactions (serious/nonserious) before and after marketing authorization	Pharmaceutical companies, regulators, international repositories	Structured, unstructured
Patient-generated data	Social media	All type of information on health from patients and physicians in online communities, facebook, twitter, etc.	Websites, blogs, and smartphone apps	Unstructured
	Monitoring devices	eHealth/mHealth monitoring devices and sensors that measure vital signs and other aspects	Monitoring devices	
Other	Scientific publications	Scientific discoveries published in scientific literature	Biomedical journals, books, etc.	Unstructured
	Epidemiological data	Various data on diseases, health, and environment	Governments, researchers, various organizations	Structured, unstructured

Notes: The main source may greatly vary depending on national health care systems and national or regional arrangements. HCP, healthcare providers; EHRs, electronic health records.

health system); data obtained from the increasing number of diagnostic tools and biomarkers available for all therapeutic areas; and in general unstructured data from e-mails, notes, text messages, paper documents, etc.

However, the simple fact that there is more data is not useful to public health unless we are able to turn it into "actionable big data" for improved health outcomes and more effective and efficient health systems. It will only inform public health decision-making if we can make meaningful extrapolations and inferences, ensuring we are quantifying the likelihood of errors to avoid offering false-positive evidence (Jordan & Mitchell, 2015). In addition to the 3Vs of big data–volume (high quantity of data), variety (very different categories of data), and velocity (fast data generation)—its veracity (quality of the data) and its value (how useful is the data) are paramount to unleash the potential for big data use in healthcare, translational science, and public health (Belle et al., 2015).

Regulators and public health decision makers have also started using novel data sources and advanced analytical methods to help them in the decision-making process or in the monitoring of approved drugs and implemented interventions (Szezák, Evers, Wang, & Pérez, 2014). In the following sections we will analyze the current and potential applications

TABLE 2 Examples of Structured and Unstructured Data

	Structured	Unstructured
Molecular and biological information	Databases with data on DNA, RNA, gene expression, proteins, structures, systems, chemical biology, ontologies, etc.	Scientific literature, annotations and linkages, written notes, conference proceedings, etc.
Clinical/Healthcare information	EHRs, databases from healthcare provider administrative information, filled forms for insurance claims, information on diagnostics, etc.	Free text in EHRs, insurance claims and medical notes
Drug development	Clinical Trial repositories, databases with clinical trial data	EPARs, biomedical journals
Safety and Pharmacovigilance	International adverse event repositories, clinical trial data, etc.	PSURs, information from social media
Environment	Weather and contamination databases, disease outbreak patterns, socioeconomic factors, lifestyle, etc.	Historical information contained in books and periodicals

Notes: EHRs, Electronic Health Records; EPAR, European public assessment reports released by the European Medicines Agency; PSURs, periodic safety update reports.

for big data in health, focusing on drug development, public health, and personal health.

2 | STRUCTURED VS. UNSTRUCTURED DATA USES

The different types of data (see Table 1) can be classified under structured, semistructured, and unstructured data (Martin-Sanchez & Verspoor, 2014). Specific examples of structured and unstructured data are listed in Table 2.

Structured data refers to information that is highly organized and can be searchable by simple and straightforward algorithms and search operations. This type of data has been used for a very long time in drug development and public health. The biggest difference nowadays is the amount of data we produce, we can store and analyze. As examples of the new applications, we have seen the advances in genomic sequencing and other "-omic" fields produce new sources of data that can help biomedical researchers unravel new hypotheses on the molecular causes of disease (Wang et al., 2013) and an increased number of data from registries and clinical records are analyzed to evaluate health interventions and their impact (Jalali, Olabode, & Bell, 2012), between others.

However, one of the new areas that are now being explored is the use of large amounts of unstructured data for drug development, drug monitoring, and public health. This is possible through the increased use of social media, the rise of information shared by patients and citizens through various platforms and mainly due to the expansion of technology capabilities that allow exploiting this data. When we refer to unstructured data, we refer to information that does not fit a predefined model or is not organized in a predefined manner (e.g., free text entries). The advantage of unstructured data is that information can be uncovered that is extremely relevant for the diagnosis and treatment planning. For example, taking a patient history might include filling out a questionnaire with a tick box whether the patient is a smoker or a nonsmoker. While the patient might tick the "nonsmoker" box, the medical doctors report on the patient history might reveal a side remark the patient made that he had just stopped smoking a couple of month ago after 25 years of smoking. Thus, unstructured information can be used to verify the veracity of structured data. Furthermore, it can contain information about risk factors that allow the prediction of readmission of patients in, e.g., congestive heart failure (Feldman, Burghard, Hanover, & Schubmehl, 2012). Social media threads and Internet searches are another form of unstructured data that is generated at high speeds, high volumes, and its veracity is often unknown. Its use and potential in public health has been widely recognized (Milinovich, Soares Magalhaes, & Hu, 2015; O'Donovan, 2015). What makes the use of unstructured data with today's technological capabilities so interesting is that its current and future purposes within and beyond the health sector can be unknown at the time of data storage. In pharmacovigilance activities, for example, data from social media could cover the deficient patient self-reporting of side effects through official channels (Broniatowski et al., 2014); which was not considered a potential use for social media data when it started being produced.

3 | USES FOR BIG DATA IN HEALTH: DRUG DEVELOPMENT

The potential use of big data in drug development lies in the discovery and development phase and in the post-marketing monitoring of drugs and devices, especially in safety monitoring and pharmacovigilance (see Figure 1). In this section, we explore these applications and also see why the use

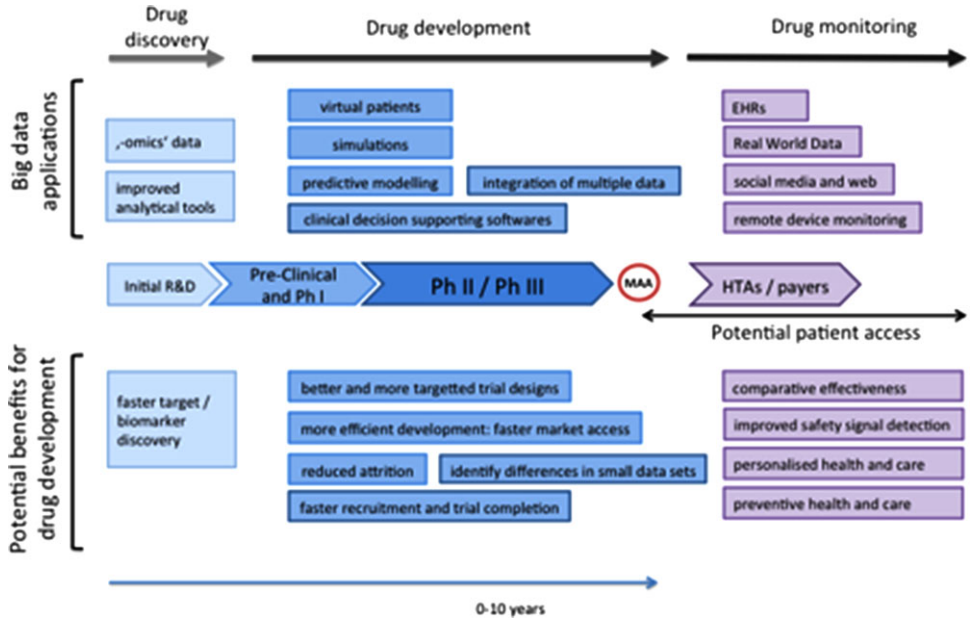

FIGURE 1 This figure illustrates the potential benefits of big data for drug development. Big data applications (top boxes) and potential benefits (bottom boxes) are divided according to the lifecycle of drug development (middle axis), from conception (drug discovery) to patient access (drug monitoring). EHRs: electronic health records; HTAs: health technology assessment bodies; MAA: marketing authorization application; Ph I/II/III: phase I/II/III clinical trials; R&D: research and development.

for regulatory purposes, outside the scope of pharmacovigilance and post-marketing activities, is currently limited or nonexistent.

3.1 | Drug discovery

Researchers nowadays have access to larger sets of data thanks to the rapidly improving analytical methods that are able to produce genomic, transcriptomic, proteomic, metabolomics, and other type of biological data faster and cheaper than ever before. Since Sanger and Coulson published their first sequencing procedure 30 years ago (Sanger & Coulson, 1975), we have developed high throughput sequencing machines that are able to map human genome sequences in less than a week and more than 22,000 times cheaper than in 2001 (see Fig. 2) (Wetterstrand, 2015). The European Bioinformatics Institute (EMBL-EBI) stores more than 20 petabytes (1 petabyte = 10^{15}bytes) of biological data and is one of the World's largest biology data repository, which can be accessed by researchers to conduct their analysis (Nature, 2013).

Integration of this wealth of "-omic" data together with the right analytical tools and the right research question or hypothesis can provide insight into associations and linkages that allow for faster drug target discovery and biomarkers for drug discovery (Szezák et al., 2014). The areas of pharmacogenomics and personalized medicine are potential candidates that will benefit from this type of big data-based research. Some public-private partnerships (PPPs) in the EU and the USA are using the big data approach for this purpose, e.g., the NIH-led PPP AMP has just launched a Big Data

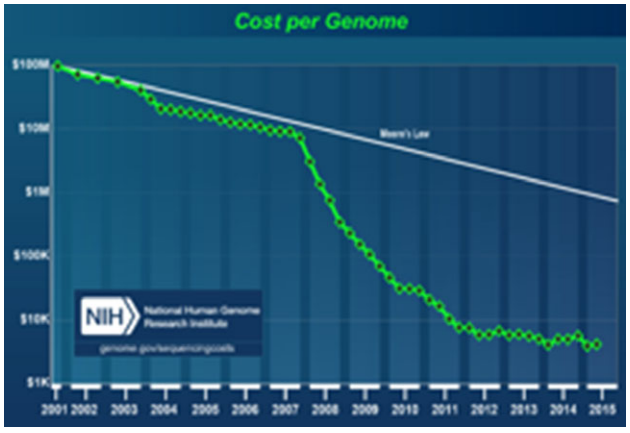

FIGURE 2 Decline in cost per genome (of human size) since 2001. From (Wetterstrand, 2015).

portal for Alzheimer's drug discovery (Sage Bionetworks, 2015). We have not seeing dramatic applications of this concept in drug development yet (Horgan et al., 2014), with the usual time to develop pharmaceuticals being close to 10 years we should analyse after 2020 whether the expectative for big data in this area has materialised.

3.2 | Drug development: R&D

The potential benefits of "rational" big data for the stakeholders involved in drug development (all stages after drug discovery) and ultimately for patients are very large. A more efficient and targeted R&D process can translate into new medicines reaching the market faster and being (in theory) cheaper. Big

data can facilitate R&D efforts in different fronts, from predictive modeling that lowers attrition to the improvement of clinical trial designs (e.g., endpoints, inclusion/exclusion criteria, etc.) that reduce trial failure and trial costs (reduced number of amendments and shorter completion times) (Raghupathi & Raghupathi 2014).

Virtual patients, created from EHRs, pharmacokinetic, and pharmacodynamic data and historical data can help power clinical trials for regulatory submission. In a recent example, the inclusion/exclusion criteria for a clinical trial assessing a new type 2 diabetes treatment were defined using information from a simulated trial that was run based on virtual patients (Garrett, O'Kelly, Walp, & Berry, 2015). Another development we have seen in the past years is the ability of clinical decision-supporting systems (CDSS), based on machine learning, to identify candidate patients for clinical trials (e.g., IBM Watson Health's Clinical Trial Matching solution) (IBM, 2014; Lehrach H, 2015). These programs have the ability to compare patients' medical and clinical information with the inclusion/exclusion criteria of recruiting clinical trials, combining information from numerous sources (e.g., EHRs, biomarkers, and international clinical trial repositories) to identify suitable candidates for participation in these trials. This can speed up the recruitment and completion of clinical trials, especially those that cover diseases with very small population pools. Various studies have evaluated the efficacy and efficiency of applying CDSS in clinical trial recruitment (Köpcke et al., 2014). Time-sensitive clinical trials in Intensive Care Units (ICU) can really profit from this technology since the usual impediment of late notification and delayed recognition hamper recruitment in these trials. Herasevich and colleagues report about an automated HER-screening process that produced alerts when eligible patients entered the ICU resulting in improved enrolment efficiency (Herasevich, Pieper, Pulido, & Gajic, 2011).

Analysis of large quantities of healthcare data (e.g., clinical trial data and EHRs) also has the potential to identify new pharmacogenomics interactions and additional indications for already existing drugs, for new drug candidates and for drugs that have failed in other indications and never reached the market. Furthermore, big data and special statistical methods can assist the development of personalised medicines, where the worldwide patient pools are very limited, by integrating data from numerous trials and facilitating the identification of differences between drugs (The Economist, 2015).

3.3 | Drug monitoring

The area of drug safety monitoring and comparative effectiveness based on "Real-World Data" is where the use of unstructured data from normal clinical practice becomes prominent and where the use of social media is being explored. Structured safety data are collected in international repositories such as VigiBase (WHO), EudraVigilance (EMA), and AERS (FDA) in the form of adverse event reports filled by drug product manufacturers (compulsory), healthcare professionals (voluntary), or patients (voluntary and still very rare). The FDA has recently opened the adverse event report database to facilitate research on safety signals and correlations; reports since 2004 are available in this web-based resource (FDA, 2014).

Safety monitoring in clinical practice is very important for all drugs and devices, but it is paramount for personalized medicines and orphan drugs (which receive marketing authorization based on data from very small patient populations) and for the new regulatory pathways that facilitate earlier patient access to medicines at the expense of higher uncertainties in the risk of these medicines (Leyens, Richer, Mellien, Ballensiefen, & Brand, 2015). Furthermore, rare adverse events may not be identified in clinical trials and only picked up through pharmacovigilance signals or real-world data.

To cover the deficiencies in patient self-reporting of adverse events, programs such as the FDA Medwatch aim to increase the structured reporting of safety information. However, new ways of identification of safety signals are needed and social media in combination with other web-resources and remote monitoring devices (or health apps) have opened a new possible pathway. The IMI project WEB-RADR (http://web-radr.eu/) launched in 2014 is evaluating how to identify potential safety signals from medicines through user comments in social media in partnership with FDA and EMA. In June 2015, another collaboration that looks into the use of patient-reported outcomes for drug safety monitoring was announced between FDA and PatientsLikeMe (PatientsLikeMe, 2015). Aspects such as duplication of information, veracity, and value will have to be carefully evaluated in such projects.

Real-world data are also being increasingly used for comparative effectiveness evaluations and postmarketing data analysis. HTA and reimbursement agencies are starting to accept outcomes data collected and processed from clinical practice in new ways, sometimes nonstandardized. Traditional postmarket studies could be complemented by information from EHRs, real-time remote monitoring devices, and data from social media (Leyens & Brand, 2016; Szezák et al., 2014).

Further regulatory uses, such as submission of big data analytics as pivotal information for marketing authorization applications, are not possible nowadays and are unlikely in the near future. At present, the regulatory applications are limited to the ones described above: to make drug development more efficient and targeted, to inform the design of pivotal studies, to monitor safety or relative effectiveness. Due to the high uncertainty in big data analysis, it can only represent supportive but not pivotal evidence.

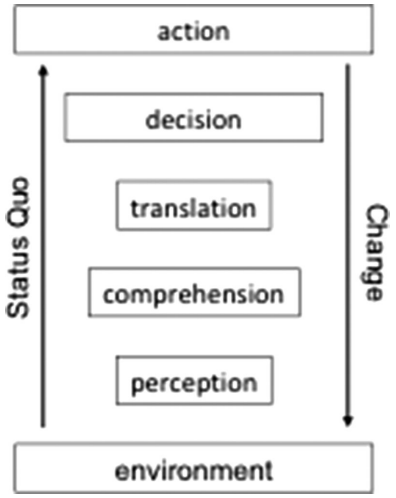

FIGURE 3 Situational awareness requires a person to perceive the environment, process the information to comprehend, and translate the data into context and knowledge so that decisions are made and actions can be taken. This process will change the status quo of the environment and situation through situational awareness and acting accordingly.

4 | USES FOR BIG DATA IN HEALTH: PUBLIC AND PERSONAL HEALTH

Public health stakeholders participate in a variety of tasks with different levels of responsibility in the health system to ensure and improve the health of the community (Ola & Sedig, 2014). The stakeholders apply analytical reasoning to the situation to make sense out of the data evidence. In general terms, public health stakeholders require great situational awareness to measure and perceive the environment at a specific given time, understand the data and translate it into information and knowledge to then form a decision and act upon the evidence base and need for action (see Fig. 3). The perfect storm of big data in public health demands for technology to make the data transparent and put it into context so that public health stakeholders can apply the cognitive task of understanding and acting upon it.

4.1 | Real-world data use to shape health policy

The aggregation of real-world data helps policy makers to guide programs and legislation. Real-world data in public health does not include just the classic health related information about an individual patient. It also comprises information about infrastructure, population, and other epidemiological factors including even weather information or socioeconomic factors. Visual analytics (Ola & Sedig, 2014) and predictive modelling (Davies et al., 2014) is an active area of research that could become an important building block to shape health policies and for public health professionals to engage in an agile manner and to predict and react to disease outbreaks (Davies et al., 2014). The role of big data in the early detection of emerging infectious diseases is becoming more and more apparent in real-world scenarios where the role of internet searches in outbreak detection (Milinovich et al., 2015) and mHealth applications, including tracking of text messages from mobile phones (O'Donovan & Bersin, 2015), become essential tools for public health organizations. In fact, telecommunication companies have started engaging in providing data for better public health monitoring and education. All information from laboratory tests to social media and environmental factors should be captured in one system that allows timely action and planning of public health measures. Davis et al. demonstrated how the use of information on disease incidence, population models, weather, and geographic information could be used in spatio-temporal epidemiologic computer simulations to predict Malaria incidence and outbreaks as well as how to test intervention strategies and their respective impact on a population level (Davies et al., 2014). Not only can these systems be used for monitoring and surveillance, making real-world data transparent and putting it into context, but they can also inform decision making processes (see Fig. 4).

4.2 | Resilient health systems

Thus, to build a sustainable, resilient health system, one must consider all sources of information captured through diverse means (e.g., mobile/social, medical professionals, wearables), integrate and analyze that data by harnessing cognitive computing to deliver the right information, insight, and knowledge for the right patient to the right care giver and medical professional. Such a resilient health systems solution can then not only integrate medical and environmental information but also data on infrastructure, resources and supplies as well as work force including skill distribution to direct the right care path for each individual patient. This concept does not only apply to developing nations, like African countries, where a clean water supply would dramatically improve the health of the population and where sustainable health systems need to be established; but also to Western nations, like the USA and European countries, where established health systems are bound to fail due to increasing costs that make health care not sustainable.

In this context, building a resilient health system that is empowered by cognitive computing and founded not only on health data but furthermore integrating environmental, geographic, and population data will create a paradigm shift in combatting diseases. The holistic view will cover the whole pipeline from data capture from a variety of sources, processing, advanced big data analytics, and predictive modeling as well as knowledge visualization and stakeholder interaction. It therefore must include the following three components:

1. End – to – end computer system for diverse data capture and integration.

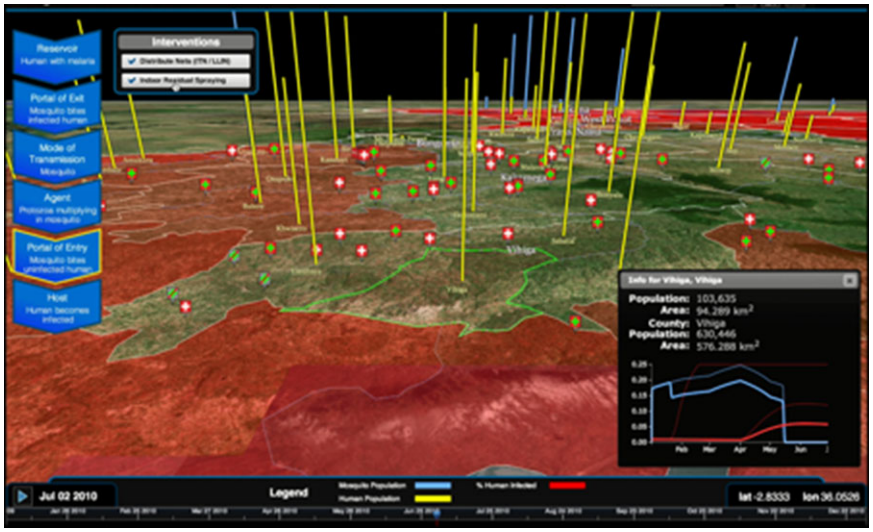

FIGURE 4 This figure shows the interface of a demonstration of IBM's Cognitive Health Care and Health systems Hub [https://www.youtube.com/watch?v = ABmIeWcmG3E]. In this scenario, malaria and interventions to reduce the disease burden are modeled using the Spatio-Temporal Epidemiological Modeler library (Davies et al., 2014). The human population and estimated mosquito count for counties in Kenya in 2010 are displayed by yellow and blue bars, respectively. On the left, the chain of causation for malaria is displayed and intervention strategies can be modeled. This illustration of malaria incidence and response has been created for demonstration purposes only and does not reflect actual data collection on malaria and intervention strategies.

2. Advanced, contextual analysis, and predictive modeling build on cognitive computing with the ability to learn to become an expert system.
3. Intuitive knowledge visualization and stakeholder interaction to intuitively deliver data, information, and knowledge in the health systems context.

The impact of such a system can be illustrated by looking at the antimicrobial resistance (AMR) challenge. It exemplifies the need to translate research and routine tests in the laboratory (bench) that leverages next-generation genome sequencing for pathogen identification to clinical practice (bedside) and to further integrate that information in infectious disease monitoring and surveillance system (society) together with the (antibiotic) treatment chosen so that public health professionals can use this information to improve treatment options and by this health outcomes and the well-being of society. Most action plans and public health initiatives to combat AMR include monitoring and tracking of pathogens, their genomes and their resistances as well as education of antibiotics use. Slowly, research actions are also supported to discover new antimicrobials and to improve the clinical and public health management of AMR. To achieve this goal, harnessing big data research and technologies will create a paradigm shift in combatting AMR by addressing the root cause of AMR for discovery and translation into actionable information and knowledge. A key concept here is to turn big data into "actionable big data" meaning that the collection and capture of data is of no consequence if the stakeholders cannot take action on it to improve the situation (see Fig. 3).

4.3 | Comparative effectiveness research and best practices

The integration of big data into a single system that allows a unified data view also facilitates clinical research to advance our understanding of disease and furthermore the best treatment and preventive options for the individual patient. In particular, comparative effectiveness research (CER) enables the comparison of benefits and risks of treating a particular set of patients with alternative methods (Zhang et al., 2013). In addition of the aforementioned use of CER in the drug development process, it is used to monitor, prevent, diagnose, and treat patients and to improve the delivery of care. The advantage of CER is that the success of an intervention is evaluated in the real-world setting vs. controlled randomized clinical trials that measure efficacy under well-controlled, ideal clinical conditions. CER is critical for providing evidence and support for informed medical care and health policy decision making on a dynamic, daily basis that adapts to the fast pace of clinical care and accelerated translation of research discoveries into clinical practice.

Visual analytics and interactive mining of data plays an important role to comprehend the information and to create hypothesis that can be tested statistically and through other scientific or differential diagnostic means. The goal is to make patient data, information, and knowledge about a disease and its treatment transparent. Clinical event patterns and pathways can be extracted from electronic health records or clinical event logs (Gotz, Wang, & Perer, 2014; Huang et al., 2014). They can then be compared with clinical pattern of other patients that underwent the same or similar treatment with a particular focus on treatment outcomes. Patient

similarity can even be extracted from heterogeneous patient records (Sun, Wang, Hu, & Edabollahi, 2012) so that integration of a variety and diverse data sources can be achieved and geared toward clinical impact. As an example, an EHR data mining algorithm was able to present questionable prescribing patterns in a cohort of hyperlipidemia patients who were occasionally given medications with side effects that have been associated in the medical literature with raising LDL levels (Perer, Wang, & Hu, 2015). As such, these tools can be used to identify best practices and improve healthcare delivery. It becomes a true win–win situation. Patients will receive more appropriate preventive care and treatment for their conditions and consequently stay healthier. This in turn has an impact on the economy of the health system that becomes less burdened with reduced costs. Medical doctors benefit directly from such a system as they can make real-time treatment decisions that are based on the best evidence available. Furthermore, policy makers can use the aggregated information to speed up decision on medical guidelines, knowledge dissemination, and funding models.

5 | CHALLENGES

Many papers and reports already discuss the challenges for the application in healthcare (European Health Forum Gastein, 2013; PerMed, 2020 CSA 2015; Science Europe, 2014), we summarize below the ones we consider crucial.

Some of these challenges are structural and inherent to a system where so many different types of data from so many varied sources come into play. All the data we store and analyze is heterogeneous, the same type of data stored in different databases may differ greatly in aspects such as accuracy, format, and detail. This data is also greatly fragmented, it is stored in numerous unconnected data sources controlled by multiple stakeholders; this is commonly referred to as information silos. It is not only the unconnected data sources that pose a challenge; the segmentation of the current biomedical research model into basic, preclinical, clinical research silos, and the healthcare model with unconnected structures for primary, secondary, tertiary, and social care also contribute to the information silos we face. Furthermore, all data is not available to all stakeholders; this may be due to intellectual protection for commercial purposes, personal privacy, cultural, and language barriers or simple lack of transparency and awareness that a certain database exists. In addition, data are not handled in a standardized way, there are no clear international codes of practice in data management, data access, data querying, or data sharing, let alone standardization in algorithms and diagnostic tests. In many instances, but especially when using social media, we have to consider the potential number of duplications of data and control for it. Most probably, data will remain heterogeneous, fragmented, unstandardized, unstructured, and unavailable; however, we can develop technological solutions to overcome most of these challenges.

Even though there have been large technological advances in the collection, storage, combination, processing, and analysis of data, there are further challenges that remain to be solved. When performing data analysis and making conclusions that inform decisions and policy making it is essential to ensure we are not offering false-positive evidence and we are quantifying the likelihood of errors. Adequate statistical methods that control for errors, such as the family-wise errors, still have to be improved and implemented as standard practice (Jordan & Mitchell, 2015). Another example is pathogenomics, where whole genome sequencing of pathogens could disrupt dramatically microbiological diagnostics. Many academic systems have shown the benefit to adopt next-generation sequencing in the determination of pathogens, their strains and resistances (Wyres et al., 2014). But it requires industry standard technology to scale out such systems across public and clinical pathology laboratories including process management, record keeping and tracking of samples, and algorithms so that the analysis can also stand in court if need be (Wyres et al., 2014).

An additional aspect that the community has to address and policy makers need to tackle today is the main ethical challenge of personal privacy. It is one of the main challenges in the use and adoption of big data for healthcare applications. The importance to preserve data sharing and openness should never come at the expense of personal privacy. In addition to the obvious precautions against hacking, all stakeholders should ensure that none of the big data uses lead to corruption and the unethical use of personal data (e.g., clinical trial data or data from EHRs). This point is especially important due to the specific nature of private medical and genetic information, which constitutes a large section of the data analyzed for health purposes. Data ownership is another ethical challenge, however it has the potential to become an opportunity if it is used and implemented in a socially acceptable manner (Hafen, Kossmann, & Brand, 2014). For big data to succeed in healthcare we need to create good policies, which are especially important to govern our "knowledge commons" and "data commons" (Frischmann, Madison, & Strandburg, 2014). The General Data Protection Regulation, currently debated at the European Council and Parliament (European Commission, 2015), will have profound implications for the use of big data in and for healthcare.

The difference in competences between stakeholders represents a further big challenge. As one example, physicians and patients will not have the same knowledge as data analysts to understand the data that is being presented to them, for this it is essential to create useful interfaces that present the data in a simple and useful manner for the specific purpose it has been analyzed. And even if the data are presented in a simple manner, some stakeholders may need additional support to interpret the data and take the right decisions based on this

data. This is especially important when presenting patients with associations between genetic data and their possible outcomes, or policy makers with epidemiological data and the possible consequences.

6 | CONCLUSION

The impact of big data in drug development, public health, and personal health and care is undeniable. Despite the challenge of heterogeneous, fragmented data that is rapidly increasing in volume at high speed; big data analytics making data transparent, easily accessible, and actionable becomes a crucial aspect in realizing the potential. Turning data into information and knowledge requires advanced analytical, cognitive computing tools, intuitive visualization, and visual analytics so that all stakeholders in drug development and public health can access the right data at the right time at the right place to take evidence–informed decisions that lead not only to an improvement in health outcomes but also to a more effective and efficient health system. It remains very clear though that the big challenges have a nature of equal variety as big data itself: Fragmented data sources, no standardization of data models and storage, analytical pitfalls in statistical methods, record keeping, and process management, patient privacy, and private interest vs. the common good.

In order to overcome these challenges we need good policies and to align the interests of all stakeholders involved in healthcare: Patients, regulators, providers, manufacturers, payers, decision makers, and researchers, among others, need to understand the need to maximize the potential of big data and "data commons" in health and follow the same ethical principles when doing so.

REFERENCES

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed Research International*, *2015*: 370194. doi: 10.1155/2015/370194. Epub 2015 Jul 2.

Bhatt, A. (2010). Evolution of clinical research: A history before and beyond james lind. *Perspective Clinical Research*, *1*(1), 6–10.

Boyle, J. (2013). Biology must develop its own big-data systems. *Nature*, *499*(7456), 7.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2014). Twitter: Big data opportunities. *Science*, *345*(6193), 148–149.

Chiolero, A. (2013). Big data in epidemiology: Too big to fail? *Epidemiology*, *24*(6), 938–939.

Davies, M., von Cavallar, S., Wyres, K. L., Reumann, M., Sepulveda, M., & Rogers, P. (2014). Intuitive information and knowledge representation of disease incidence and respective intervention strategies. *Studies in Health Technology and Informatics*, *205*, 1173–1177. doi 10.3233/978-1-61499-432-9-1173

European Commission. (2015). http://ec.europa.eu. Retrieved from http://ec.europa.eu/justice/data-protection/review/index_en.htm (accessed August 12, 2015).

European Health Forum Gastein. Big data workshop. Workshop report, Gastein: European Health Forum Gastein, 2013.

FDA. (2014). *open.fda.gov*. Retrieved from https://open.fda.gov/drug/event/ (accessed August 12, 2015).

Feldman, S, Burghard, C., Hanover, J, & Schubmehl, D. (2012). Unlocking the power of unstructured data. *IDC Health Insights*, *#HI235064*, 1–10.

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (2014). *Governing knowledge commons*. Oxford: Oxford University Press.

Garrett, A., O'Kelly, M., Walp, D., & Berry, N. S. (2015). Lifecycle modeling and simulation in clinical trials. *Applied Clinical Trials*, *24*(6).

Gotz, D., Wang, F., & Perer, A. (2014). A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, *48*, 148–159.

Hafen, E., Kossmann, D., & Brand, A. (2014). Health data cooperatives—citizen empowerment. *Methods of Information in Medicine*, *53*(2), 82–86.

Herasevich, V., Pieper, M. S., Pulido, J., & Gajic, O. (2011). Enrollment into a time sensitive clinical study in the critical care setting: Results from computerized septic shock sniffer implementation. *Journal of the American Medical Informatics Association*, *18*(5), 639–644.

Horgan, D., Jansen, M., Leyens, L., Lal, J. A., Sudbrack, R., Hackenitz, E., … Brand, A. (2014). An index of barriers for the implementation of personalised medicine and pharmacogenomics in Europe. *Public Health Genomics*, *17*, 287–298.

Huang, Z., Dong, W., Ji, L., Gan, C., Lu, X., & Duan, H. (2014). Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, *47*, 39–57.

IBM. (2014). Mayo clinic and IBM task watson to improve clinical trial research. http://www-03.ibm.com/. Retrieved from http://www-03.ibm.com/press/us/en/pressrelease/44754.wss (accessed August 12, 2015).

Jalali, A, Olabode, O. A., & Bell, C. M. (2012). Leveraging cloud computing to address public health disparities: An analysis of the SPHPS. *Online Journal of Public Health Information*, *4*(3): ojphi.v4i3.4325.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Köpcke, F., & Prokosch, H-U. (2014). Employing computers for the recruitment into clinical trials: A comprehensive systematic review. *Journal of Medical Internet Research*, *16*(7), e161.

Lehrach, H. (2015). Virtual clinical trials, an essential step in increasing the effectiveness of the drug development process. *Public Health Genomics*, *18*(6), 366–371

Leyens, L., & Brand, A. (2016). Early patient access to medicines: Health technology assessment bodies need to catch up with new marketing authorization methods. *Public Health Genomics*, *19*, 187–191.

Leyens, L., Richer, E., Mellien, Ø., Ballensiefen, W., & Brand, A. (2015). Available tools to facilitate early patient access to medicines in the EU and the USA: Analysis of conditional approvals and the implications for personalized medicine. *Publich Health Genomics*, *18*(5), 249–259.

Martin-sanchez, F., & Verspoor, K. (2014). Big data in medicine is driving big change. *Yearbook of medical informatics*, *9*, 14–20.

Milinovich, G. J., Soares Magalhaes, R. J., & Hu, W. (2015). Role of big data in the early detection of ebola and other emerging infectious diseases. *Lancet*, *3*, e20–e21.

Nature. (2013). The big challenges of big data. *Nature*, *498*, 255.

O'Donnovan, J., & Bersin, A. (2015). Controlling Ebola through mHealth strategies. *Lancet*, *3*, e22.

Ola, O., & Sedig, K. (2014). The challenge of big data in public health: An opportunity for visual analytics. *Online Journal of Public Health Informatics*, *5*(3), e223.

PatientsLikeMe. (2015). http://blog.patientslikeme.com. *PatientsLikeMe and the FDA Sign Research Collaboration Agreement*. Retrieved from http://blog.patientslikeme.com/2015/06/15/patientslikeme-and-the-fda-sign-research-collaboration-agreement/ (accessed August 12, 2015).

Perer, A., Wang, F., & Hu, J. (2015). Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, *56*, 369–378.

PerMed 2020, CSA. (2015). *The PerMed SRIA: 'Shaping Europe's vision for personalised medicine*. Koeln (Cologne): German Aerospace Center (DLR).

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*, 3.

Sage Bionetworks. (2015). *www.synapse.org*. Retrieved from https://www.synapse.org/#!Synapse:sYn2580853/ (accessed August 12, 2015).

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448.

Science Europe. (2014). *How to transform big data into better health: envisioning a health big data ecosystem for advancing biomedical research and improving health outcomes in Europe*. Workshop report, Brussels: Science Europe.

Sun, J., Wang, F., Hu, J., & Edabollahi, S. (2012). ACM. *SIGKDD Explorations*, *14*(1), 16–24.

Szezák, N., Evers, M., Wang, J., & Pérez, L. (2014). The role of big data and advanced analytics in drug discovery, development, and commercialization. *Clinical Pharmacology and Therapeutics*, *95*(5), 492–495.

The Economist. (2015). Medicine by numbers. *The Economics Technology Quarterly*, Mar 7th, 2015, 19–20.

Wang, L. W., Qu, A. P., Yuan, J. P., Chen, C., Sun, S. R., & Hu, M. B. (2013). Computer-based image studies on tumor nests mathematical features of breast cancer and their clinical prognostic value. *PLoS One*, *8*(12), e82314.

Wetterstrand, K. A. (2015). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). Retrieved from . http://www.genome.gov/sequencingcosts/ (accessed August 12, 2015).

Wyres, K. L., Conway, T. C., Garg, S., Queiroz, C., Reumann, M., Hogg, G., … Rusu, L. I. (2014).WGS analysis and interpretation in clinical and public health microbiology laboratories: What are the requirements and how do existing tools compare? *Pathogens*, *3*(2), 437–458.

Zhang, S., Li, L., Yu, Y., Sun, X., Xu, L., Zhao, W., … Pan, Y. (2013). A novel approach to generating CER hypotheses based on mining clinical data. *Studies in Health Technology and Informatics*, *192*, 991.