Data Science in Healthcare: Benefits, Challenges and Opportunities

Ziawasch Abedjan, Nozha Boujemaa, Stuart Campbell, Patricia Casla, Supriyo Chatterjea, Sergio Consoli, Cristobal Costa-Soria, Paul Czech, Marija Despenic, Chiara Garattini, Dirk Hamelinck, Adrienne Heinrich, Wessel Kraaij, Jacek Kustra, Aizea Lojo, Marga Martin Sanchez, Miguel A. Mayer, Matteo Melideo, Ernestina Menasalvas, Frank Moller Aarestrup, Elvira Narro Artigot, Milan Petković, Diego Reforgiato Recupero, Alejandro Rodriguez Gonzalez, Gisele Roesems Kerremans, Roland Roller, Mario Romao, Stefan Ruping, Felix Sasaki, Wouter Spek, Nenad Stojanovic, Jack Thoms, Andrejs Vasiljevs, Wilfried Verachtert, and Roel Wuyts

Authors are listed in alphabetic order since their contributions have been equally distributed.

Z. Abedjan · R. Roller · J. Thoms
DFKI GmbH, Berlin, Germany

N. Boujemaa
Inria Saclay Ile-de-France, Paris, France

S. Campbell
Information Catalyst, Northwich, UK

P. Casla · A. Lojo
IK4-IKERLAN, Arrasate-Mondragon, Spain

S. Chatterjea · S. Consoli (✉) · M. Despenic · A. Heinrich · J. Kustra · M. Petković
Philips Research, Eindhoven, The Netherlands
e-mail: sergio.consoli@philips.com

C. Costa-Soria
Intituto Tencologico de Informatica (ITI), Valencia, Spain

P. Czech
Know-Center GmbH, Graz, Austria

C. Garattini · M. Romao
Intel Corporation NV/SA, Kontich, Belgium

D. Hamelinck · W. Verachtert · R. Wuyts
IMEC, Leuven, Belgium

W. Kraaij
TNO, The Hague, The Netherlands
Leiden University, Leiden, The Netherlands

S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_1

1 Introduction and Preliminaries

An improvement in health leads to economic growth through long-term gains in human and physical capital, which ultimately raises productivity and per capita GDP [27, 35, 61]. The healthcare sector currently accounts for 10% of the EU's GDP. In 2014 the EU-28's **total healthcare expenditure was € 1.39 trillion**. This is expected to increase to 30% by 2060. The increase in healthcare costs is primarily due to a rapidly ageing population (e.g. proportion of individuals aged 65 years and older is projected to grow from 15% in 2000 to 23.5% by 2030), rising prevalence of chronic diseases and costly developments in medical technology. Chronic diseases result in the **loss of 3.4 million potential productive life years**. This amounts to an **annual loss of € 115 billion** for EU economies. However, the EU spends **only 3%** of its healthcare budget on prevention, with chronic diseases being among the most preventable illnesses (https://euobserver.com/chronic-diseases/125922).

M. M. Sanchez
Huawei Technologies, Munich, Germany

M. A. Mayer
Universitat Pompeu Fabra, Barcelona, Spain

M. Melideo
Engineering Ingegneria Informatica SPA, Roma, Italy

E. Menasalvas · A. R. Gonzalez
Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain

F. M. Aarestrup
Technical University of Denmark, Lyngby, Denmark

E. N. Artigot
Everis, Centro Empresarial el Trovador, Zaragoza, Spain

D. Reforgiato Recupero
University of Cagliari, Cagliari, Italy

G. R. Kerremans
European Commission, Luxembourg City, Luxembourg

S. Ruping
Fraunhofer-Institut fur Intelligente Analyse, Sankt Augustin, Germany

F. Sasaki
Cornelsen GmbH, Berlin, Germany

W. Spek
T.I.B. Development, Vlaardingen, The Netherlands

N. Stojanovic
Nissatech Innovation Centre, Nis, Serbia

A. Vasiljevs
Tilde, Riga, Latvia

The relatively large share of public healthcare spending in total government expenditure underscores the need to improve the sustainability of current health system models. However, the effectiveness of a healthcare system depends on three components, namely, *quality*, *access* and *cost*. To improve productivity of the healthcare sector, it is necessary to reduce cost *while* maintaining or improving the quality of care provided. The fastest, least costly and most effective way to achieve this is to use the knowledge that is hiding within the *already existing* large amounts of generated medical data (http://www.healthparliament.eu/documents/10184/0/EHP_papers_BIGDATAINHEALTHCARE.pdf/8c3fa388-b870-47b9-b489-d4d3e8c64bad). According to current estimates, medical data is already in the zettabyte scale and will soon reach the yottabyte (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/). While most of this data was previously stored in a hard copy format, the current trend is towards digitization of these large amounts of data resulting in what is known as *Big Data*.

This chapter provides an overview of needs, opportunities and challenges of using (Big) Data Science technologies in the healthcare sector, including several recommendations:

- **Breaking down data silos in healthcare.** Access to high-quality, large healthcare datasets will optimize care processes, disease diagnosis, personalized care and in general the healthcare system. Furthermore, true transformation of the healthcare sector can only be achieved if all stakeholders and verticals in the healthcare sector (healthtech industry, healthcare providers, pharma and insurance companies, etc.) share big data and allow free data flow.
- **Standardization and interoperability.** In the healthcare sector, data is often fragmented or generated in different systems with incompatible formats. Therefore, interoperability and standardization are key to deploy the full potential of data.
- **Privacy and ethics.** Health data presents specific challenges and opportunities. Better clinical outcomes, more tailored therapeutic responses and disease management with improved quality of life are all appealing aspects of data usage in health. However, because of the personal and sensitive nature of health data, special attention needs to be paid to legal and ethical aspects concerning privacy, as well as to privacy-preserving technologies that can overcome these barriers.
- **Increased focus on prevention.** Currently, 97% of healthcare budgets are spent on treating patients both with acute and chronic conditions. Only 3% is spent on prevention, with chronic diseases being among the most preventable illnesses. Considering the economic impact of chronic diseases on the productivity of the EU workforce, an increased focus on primary and secondary prevention is clearly needed.

- **Policy.** Dealing with different health data protection regimes across EU Member States creates difficulties in accessing and sharing health data at EU level. The implementation of the GDPR is an opportunity to look for alignment. Finally, innovative approaches to healthcare, such as value-based healthcare, should be supported by policy to drive the transformation of the healthcare sector. Developing policies and technologies will contribute towards enabling the digital single market strategy.

To prove the impact of these recommendations, it is essential to demonstrate the value created by Data Science in large-scale pilots. These pilots are meant to serve as the best practice examples of transforming the health sector with the aim to increase its quality, decrease costs and improve accessibility. This can be done by putting Data Science technologies at their core with the goal that their results can be scaled up and potentially transferred to other sectors.

2 Healthcare Opportunities

The healthcare [35] sector currently accounts for 8% of the total European workforce and for 10% of the EU's GDP [31]. However, **public expenditure on healthcare and long-term care is expected to increase by one third by 2060** [35]. This is primarily due to a rapidly ageing population, rising prevalence of chronic diseases and costly developments in medical technology. The relatively large share of public healthcare spending in total government expenditure, combined with the need to consolidate government budget balances across the EU, underscores the need to improve the sustainability of current health system models. Evidence suggests that **by improving the productivity of the healthcare system, public spending savings would be large, approaching 2% of GDP on average in the OECD** [30] which would be equivalent to **€ 330 billion in Europe** based on GDP figures for 2014 [27].

Data Science technologies have already made some impact in fields related to healthcare: medical diagnosis from imaging data in medicine, quantifying lifestyle data in the fitness industry, just to mention a few. Nevertheless, for several reasons that will be discussed in the book, healthcare has been lagging in taking data analytics approaches, which is a paradoxical situation, since it was already estimated by the Ponemon Institute in 2012 that 30% of all the electronic data storage in the world was occupied by the healthcare industry [29]. It is evident that within existing mounds of big data, there is hidden knowledge that could change the life of a patient or, at a very large extent, change the world itself. **Extracting this knowledge is the fastest, least costly and most effective path to improving people's health** (http://www.healthparliament.eu/documents/10184/0/EHP_papers_BIGDATAINHEALTHCARE.pdf/8c3fa388-b870-47b9-b489-d4d3e8c64bad).

Data Science technologies will definitely open new opportunities and enable breakthroughs related to, among the others, healthcare data analytics (http://www.gartner.com/it-glossary/predictive-analytics/) addressing different perspectives: (1) **descriptive**, to answer what happened; (2) **diagnostic**, to answer the reason why it happened; (3) **predictive**, to understand what will happen; and (4) **prescriptive**, to detect how we can make it happen.

It is out of any doubt that the potential impact of Data Science on technology, economic and society is extremely relevant, boosting innovations in organizations and leading to the improvement of business models. This chapter emphasizes that Data Science has the potential to unlock vast productivity bottlenecks and radically improve the quality and accessibility of the healthcare system and discusses steps that need to be taken towards a large and in-depth adoption.

2.1 Economic Potential

The rapidly ageing population is contributing to the ever-increasing demands as chronic diseases are more prevalent in the elderly. The number of people aged 85 years and older is projected to rise from 14 million to 19 million by 2020 and to 40 million by 2050 [32]. The effect of these ever-increasing demands is clearly illustrated by a study conducted by Accenture in 2014 which found that a third of European hospitals had reported operating losses [1]. This only exacerbates the fact that countries in Europe are finding it increasingly challenging to provide good-quality care at a reasonable cost to their citizens when it is needed [61]. The concept of the **Iron Triangle of Healthcare** [38] is often quoted to describe this very challenge. The three components of the triangle are **quality, access and cost**. Efficacy, value and outcome of the care reflect the quality of a healthcare system. Access describes who can receive care when they need it. Cost represents the price tag of the care and the affordability of the patients and payers. The problem is that all the components are typically in competition with one another in the healthcare sector. Thus while it may be possible to improve any one or two components, in most of the cases this comes at the expense of the third [38], as illustrated in Fig. 1.

However, while the present healthcare optimization approaches may help introduce minor changes in the balance of the Iron Triangle of Healthcare, only a radical breakthrough has the potential to totally disrupt the Iron Triangle of Healthcare such that all three components including quality, access and cost are all further optimized simultaneously. Given that healthcare is one of the most data-intensive industries around, the multitude of high volume, high variety, high veracity and value of data sources within the healthcare sector has the potential to disrupt the Iron Triangle of Healthcare. While most of this healthcare data was previously stored in a hard copy format, the current trend is towards digitization of these large amounts of data, which can facilitate this process.

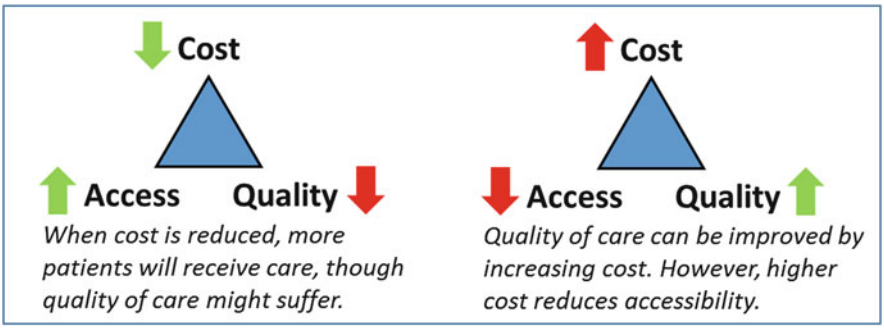

Fig. 1 The examples indicate how current approaches to healthcare improvement often lead to suboptimal solutions

2.2 Technical and Organizational Challenges

Although there is already a huge amount of healthcare data around the world and while it is growing at an exponential rate, nearly all of the data is stored in individual silos [14]. Data collected by a general practice (GP) clinic or by a hospital is mostly kept within the boundaries of the healthcare provider. Moreover, data stored within a hospital is hardly ever integrated across multiple IT systems. For example, if we consider all the available data at a hospital from a single patient's perspective, information about the patient will exist in the EMR system, laboratory, imaging system and prescription databases. Information describing which doctors and nurses attended to the specific patient will also exist. However, in the vast majority of cases, every data source mentioned here is stored in separate silos. Thus deriving insights and therefore value from the aggregation of these datasets is often not possible at this stage. It is also important to realize that in today's world a patient's medical data does not only reside within the boundaries of a healthcare provider. The medical insurance and pharmaceutical industries also hold information about specific claims and the characteristics of prescribed drugs, respectively. Increasingly, patient-generated data from IoT (Internet of Things) devices such as fitness trackers, blood pressure monitors and weighing scales provide critical information about the day-to-day lifestyle characteristics of an individual. Insights derived from such data generated by the linking among EMR data, vital data, laboratory data, medication information, symptoms (to mention some of these) and their aggregation, even more with doctor notes, patient discharge letters, patient diaries and medical publications, namely, linking structured with unstructured data, can be crucial to design coaching programmes that would help improving peoples' lifestyles and eventually reduce incidences of chronic disease, medication and hospitalization.

As the healthcare sector transitions from a volume- to value-based care model, it is essential for different stakeholders to get a complete and accurate understanding of treatment trajectories of specific patient populations. The only way to achieve this is to be able to aggregate the disparate data sources not just within a single hospital's

IT infrastructure but also across multiple healthcare providers, other healthcare players (e.g. insurance and pharma) and even consumer-generated data. Such unified datasets would not only bring benefits to every player within the healthcare industry (thus allowing better-quality care and access to healthcare at lower costs) but the population health in general, and the patient in particular, by providing first-time right treatment based on a sustainable pricing model.

However, achieving such a vision which involves the integration of such disparate healthcare datasets in terms of data granularity, quality and type (e.g. ranging from free text, images, (streaming) sensor data to structured datasets) poses major legal, business and technical challenges from a data perspective, in terms of the volume, variety, veracity and velocity of the datasets. The only way to successfully address these challenges is to utilize big data and Data Science.

"Big Data" has a wide range of definitions in health research [5, 51]. However, a viable definition of what Big Data means for healthcare is the following: "Big Data in Health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points" [4]. A more general definition of Big Data refers to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse" (McKinsey Global Institute). This definition puts the accent on size/volume, but, as we stated above, the dimensions are many: variety (handling with a multiplicity of types, sources and format), data veracity (related to the quality and validity of these data) and data velocity (availability in real time). In addition, there are other factors that should also be considered such as data trustworthiness, data protection and privacy (due to the sensitivity of data managed). All these aspects lead to the need for new algorithms, techniques and approaches to handle these new challenges.

3 Opportunities with Most Impact

This section describes particular areas in health (including healthy living and healthcare) that would most benefit from the application of Data Science.

3.1 Healthy Living: Prevention and Health Promotion

3.1.1 Lifestyle Support

Data analytics technologies could help provide more effective tools for behavioural change. Especially mobile health (mHealth) has the potential to personalize interventions, taking advantage of lifestyle data (nutrition, physical activity, sleep) and coaching style effectiveness data from large reference populations. Besides providing information to people, mHealth technologies exploit contextual information

which is the key to personal and precision medicine. This can help provide a fully integrated picture of what influences progress and setbacks in therapy.

3.1.2 Better Understanding of Triggers of Chronic Diseases for Effective Early Detection

Data Science tools can support ongoing research into better understanding the relation between social and physical behaviours, nutrition, genetic factors, environmental factors and the development of mental/physical diseases. The complex interactions between the different systems that determine disease progression are still not fully understood, and it is expected that an integrated view of health based on various markers (i.e. omics, quantified self-data) can help improve early detection of diseases and long-term management of adverse health factors, thereby reducing costs.

3.1.3 Population Health

Public health policy is based on a thorough analysis of the health status of a population stratified by region and socio-economic status (SES) in order to define and focus on societal actions to improve health outcomes. Big data analysis can guide policies to address a certain population segment by specific interventions. The success of the policy is critically dependent on the quality of the underlying research and the quality (effectiveness) of the interventions. For many interventions (for instance, in the social/mental health domain), universally accepted methods for validating success are still lacking. There are several challenges regarding Data Science and population health such as:

- Data protection regulation makes it difficult to analyse data from different healthcare providers and services in combination;
- A significant part of the population health records is unstructured text;
- There are interoperability, data quality and data integration limitations;
- Existing systems are not dynamically scalable to manage and maintain Big Data structures.

The large-scale, systematic and privacy-respecting measurement and collection of outcomes along with careful validation involving advanced statistical methods for handling missing data will allow for strengthening the evidence base for policymaking and developing more precise and effective (stratified/personalized) interventions.

3.1.4 Infectious Diseases

Technology in recent years has made it possible to not only get data from the healthcare environment (hospitals, health centres, laboratories, etc.) but also information from society itself (sensors, monitoring, IoT devices, social networks, etc.). The health environments would benefit directly through the acquisition and analysis of the information generated in any kind of social environment such as social networks, forums, chats, social sensors, IoT devices, surveillance systems, virtual worlds, to name a few. These environments provide an incredible and rich amount of information that could be analysed and applied to the benefit of public health. Combining information from informal (e.g. web-based searches and Google) and syndromic surveillance and diagnostic data including the next-generation sequencing can provide much earlier detection of disease outbreaks and detailed information for understanding links and transmission [9]. The ARGO [39] model, for instance, uses several data sources, including Google search data to create a predictive model for influenza. Different systems have been created to track disease activity levels (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467) or spread dynamics and surveillance (http://dl.acm.org/citation.cfm?id=2487709; http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055205; http://link.springer.com/article/10.1007/s10916-016-0545-y) using social information provided by Twitter. Analysing these data in combination with explanatory variables, such as travel, trade, climate changes, etc., could allow for the development of predictive models for population-based interventions as well as improved individual patient treatment. Governmental public health experts can better detect early signs of disease outbreaks (http://searchhealthit.techtarget.com/feature/Social-data-a-new-source-for-disease-surveillance; e.g. influenza, bacterial-caused food poisoning) and coordinate quarantine and vaccination responses.

3.2 Healthcare

3.2.1 Precision Medicine

The systematic collection and analysis of genetic data in combination with diseases, therapies, and outcomes has the potential to dramatically improve the selection of the best treatments, avoiding the harming of patients, and the use of ineffective therapies. The availability of historical longitudinal patient data concerning environmental exposure and lifestyle would also help better determine the (ensemble of) causes triggering the onset of a disease state. An important new technology driving precision medicine is high-performance genome analysis. The vast amount of genomic data that will become available enables new analytical algorithms for clinical use. It will, for example, become possible to compare whole genomes of patients against a large population of other individuals. Screening large genomic databases for rare diseases located at different centres is such an example. This

process is complex, since data is non-centralized, and—if the data is not readily available yet—it requires large amounts of computing power.

3.2.2 Collecting Patient-Reported Outcomes and Total Pathway Costs for Value-Based Healthcare

A guiding principle for sustainable healthcare is "value-based healthcare" (VBHC) Porter [62], where patient-reported outcomes, normalized by the total cost of the care path, determine the decision to pay for a specific treatment. In healthcare, pay for performance is a model that offers financial incentives to healthcare provider for improving quality and effectiveness of healthcare by meeting certain performance measures (e.g. a healthcare provider is not paid for the time spent treating a patient but for the outcome). In order to make VBHC reality, data must be collected, analysed and aggregated regarding care paths, therapies and costs. In particular, "patient-related health outcomes" need to be collected and verified before, during and after treatments, all of which is not currently common practice. On the other hand, it is also a challenge to reorganize administrative care systems to be able to connect all the involved costs of specific care paths in order to have an accurate estimate of the full costs involved. As soon as care processes have been linked and care paths can be traced, decisions for particular therapies can be based on empirical evidence, as supported by a huge database of "patient-reported outcomes" (patient self-assessment of health parameters based on, e.g. questionnaires and tracking devices) of patients with similar diseases and the associated total cost of treatments and therapies. It is essential that the methods to collect patient-reported health outcomes and costs per therapy/care path are standardized and validated.

3.2.3 Optimizing Workflows in Healthcare

The manufacturing industry involves processes which are in many cases predictable. However, conditions within a hospital are highly dynamic and often dependent on a huge number of interrelated factors spanning the patients themselves and their needs, multiple departments, staff members and assets. This volatile situation makes any form of workflow orchestration to improve productivity highly challenging unless hospital staff and administrators have a proper overview of the hospital's operation. This makes it essential for a healthcare provider to have necessary tools to integrate multiple data streams such as real-time location tracking systems, electronic medical records, nursing information systems, patient monitors, laboratory data and machine logs to automatically identify the current operational state of a hospital. This allows more effective decision-making that results in better resource utilization and thus higher productivity and quality.

3.2.4 Infection Prevention, Prediction and Control

Data Science can make a difference in very specific healthcare challenges too. For example, infection control is the discipline concerned with preventing hospital-acquired or healthcare-associated infection (HAI). According to the European Centre for Disease Prevention and Control [22], 100,000 patients are estimated to acquire a healthcare-associated infection in the EU each year. The number of deaths occurring as a direct consequence of these infections is estimated to be at least 37,000, and these infections are thought to contribute to an additional 110,000 deaths each year. It is estimated that approximately 20–30% of healthcare-associated infections are preventable by intensive hygiene and control programmes. Furthermore, the Centres for Disease Control and Prevention in the USA estimated 722,000 HAIs in US acute care hospitals in 2011. About 75,000 hospital patients with HAIs died during their hospitalizations [26]. Preventing HAIs could save $25–32 billion in the USA alone [58]. The World Health Organization has strict guidelines on protocols that need to be followed to minimize the risk of the spread of infection. While some of the guidelines are easy to implement and follow, there are others that are hard to implement simply due to the lack of any technology that can ensure strict adherence to the guidelines. Real-time and big data technologies are needed to integrate genomics with epidemiology data not to just control but also prevent and predict the spread of infections within a healthcare setting.

3.2.5 Social-Clinical Care Path

Healthcare is moving towards an integrated care approach, which according to the definition of the World Health Organization (WHO) is "a concept bringing together inputs, delivery, management and organization of services related to diagnosis, treatment, care, rehabilitation and health promotion. Integration is a means to improve services in relation to access, quality, user satisfaction and efficiency [24]". Care integration means the involvement of both clinical and social actors (e.g. care workers) which are active in care management after the patients are discharged from the hospital but still need assistance and care. This defines new pathways involving different actors from different domains all managing and generating data evolving around the patient. The data collected in the operation of these care pathways can be used to identify inefficiencies and to recommend "optimal treatment pathways" [43].

3.2.6 Patient Support and Involvement

In addition to collecting patient-reported health outcomes, there are other opportunities for patient empowerment and involvement. Notable examples are patient-centred care paths, patient-controlled health data and shared decision-making of clinicians together with patients. For all these methods, the control of patients on their own health data is vital. The patient controls for managing health

data should support different levels of digital/health literacy and allow tracking patient consent of opting in/out for clinical research studies. For example, web fora of patient organizations play an important role in exchanging information about disease, medication and coping strategies, complementary to regular patient briefing information. Recent studies show that mining these fora can yield valuable hypotheses for clinical research and practice (e.g. chronomedication or side effects [41]). Also, new approaches to interact with the general population directly, e.g. via crowdsourcing, analysing search logs (http://blogs.microsoft.com/next/2016/06/07/how-web-search-data-might-help-diagnose-serious-illness-earlier/#sm.0001mr81jwowvcp6zs81tmj7zmo81) or AI-based chatbots, are ways to collect information that previously was not available.

3.2.7 Shared Decision Support

By emphasizing the patient's involvement within decision processes, patients are able to gain a better understanding of all health-related issues. In this sense, giving patients control over and insight in their own health data can help strengthen patient-centred care after decades of a disease-centred model of care and allow the easier customization of healthcare and precision medicine. Logically, lifestyle data collected and aggregated into meaningful information should motivate patients to achieve higher compliance rates and lower pharmaceutical costs. Meaningful information critically depends on the ability of systems to quantify the inherent uncertainty involved in the diagnosis and also the uncertainty with respect to the outcomes of treatment alternatives and associated risks.

3.2.8 Home Care

Professional tracking and recording of medical data as well as personal data should not be limited to only hospitals and doctors. Due to demographic changes, new models for home care or outpatient care (facilities) have to be developed. Data Science can support the general ICT-based transformation in this area. By combining smart home technologies, wearables, clinical data and periodic vital sign measurements, home care providers could remotely support, by an expanded healthcare infrastructure, individuals (chronically ill or elderly), who will be empowered to live longer on their own.

3.2.9 Clinical Research

The integration and analysis of the huge volume of health data coming from many different resources such as electronic health records, social media environments, drug and toxicology databases and all the "omics" data such as genomics, proteomics and metabolomics is a key driver for the change from (population-level)

evidence-based medicine towards precision medicine. Data Science can enhance clinical research by:

- discovering hidden patterns and associations within the heterogeneous data, uncovering new biomarkers and drug targets
- allowing the development of predictive disease progression models;
- analysing real-world data (RWD) as a complementary instrument to clinical trials, for the rapid development of new personalized medicines (http://www.pmlive.com/pharma_thought_leadership/the_importance_of_real-world_data_to_the_pharma_industry_740092). The development of advanced statistical methods for learning causal relations from large-scale observational data is a crucial element for this analysis.

A prerequisite for the effective use and reuse of the various kinds of data for clinical research is that the data is FAIR (F*indable*, A*ccessible*, I*nteroperable*, R*eusable*) [63]. To support this requirement, organizations like the World Wide Web Consortium (W3C) have worked on the development of interoperability guidelines (https://www.w3.org/blog/hcls/) in the realm of healthcare and life sciences.

3.3 Healthcare Data Stewardship Challenges

In addition to requiring data to be FAIR, it is also crucial to store health data in secure and privacy-respecting databases. Trustworthiness is the main concern of individuals (citizens and patients) when faced with the usage of their health-related data. Intentional or unintentional disclosure of, e.g. medication record, lifestyle data and health risks can compromise individuals and their relatives. National governments and the EU are faced with the problem of integrating the diverse legal regulations and practices on sensitive data and their analysis. This has to fit to the needs of society (all of society, including patients), research institutes, medical institutes, insurance schemes and all healthcare providers, as well as companies and many more stakeholders.

Currently various approaches exist for analysing data sources available in a specific domain or for connecting these different databases across domains or repositories. Still several conflicts and risks have to be addressed to accomplish the ambitious plan of combining health databases by new anonymization and pseudonymization approaches to guarantee privacy. Analysis techniques need to be adapted to work with encrypted or distributed data [50]. The close collaboration between domain experts and data analysts along all steps of the data analytics chain is of utmost importance.

4 Privacy, Ethics and Security

This section will document the regulations, which influence and drive the adoption of Data Science in terms of privacy, data protection and ethics.

In this increasingly digital and connected world, where there are more opportunities to access and combine databases from various sources, we can assume that more insights and information can and will be derived from records of patient data/people's activities. This implies that various parties could also misuse the new discovery [28]. In this respect, a lot of skepticism with regard to "where the data goes to", "by whom it is used" and "for what purpose" is present in most public opinion, and, so far, European and international fragmented approaches together with an overly complex legal environment did not help.

However, a new General Data Protection Regulation (GDPR), replacing the previous Data Protection Directive (1995), was adopted in April 2016 and aims at harmonizing legislation across EU Member States. As a "regulation", the GDPR applies to all Member States without the need of transposition into national legislation. The GDPR was implemented by mid-2018 to allow public and private sector to adapt their organizational measures to the new legal framework.

The Regulation also provides a margin of manoeuvre for Member States to specify their rules including the processing of special categories of personal data ("sensitive data"). Thus the Regulation does not prevent Member States' law from setting out the circumstances for specific processing situations, e.g. introducing "further conditions, including limitations, with regard to the processing of genetic data, biometric data, and data concerning health". As a result, it is probable that different data protection implementations for health data will continue persisting across the European Union. To enable the single EU digital market also in the healthcare sector, it is of utmost importance to harmonize the national member state laws that regulate sensitive health data.

The adopted legislation went through long discussions and reflects a tension between fostering and facilitating innovation (e.g. establishment of a single European Data Protection Board comprising all national data protection authorities, harmonization of laws, etc.) and a political drive to protect privacy and enable individual citizens' control over their data. The latter is strictly connected with Articles 7 and 8 in the Charter of Fundamental Rights of the European Union on the "respect for private and family life" and the "protection of personal data", respectively.

Health data presents specific challenges and opportunities. Better clinical outcomes, more tailored therapeutic responses and disease management with improved quality of life are all appealing aspects of data usage in health. However, because of the personal and sensitive nature of health data, special attention needs to be paid to legal and ethical aspects concerning privacy. To unlock its potential, health (and genomic) data sharing, with all the challenges it presents, is often necessary, and much work is currently being done to ensure such endeavours are undertaken responsibly (https://genomicsandhealth.org/about-the-global-alliance/

key-documents/framework-responsible-sharing-genomic-and-health-related-data). In this context, the temptation needs to be resisted to see free data flow and data protection as irreconcilable opposites.[1] Data sharing can bring benefit at individual and societal levels and therefore should be further promoted; for example, organizations can put in place appropriate technical and organizational measures to mitigate privacy risks.

Besides top-down approaches to protect the privacy of people, there are other ways in which the community can enhance ethical approaches to data and support the understanding of the delicate nuances of working in this field. Internet data and big data tend to blur the lines between areas that are traditionally perceived as separate and that are a stronghold of how to use data and, for example, do research on these. They complicate the distinction between what is public and private (e.g. social media), between people and the data they produce, whether data producers can be considered "human subjects" for research and if people are even aware of being such a subject (e.g. passive sensing) and finally raise issues on accountability, transparency and the unanticipated consequences of automation (e.g. algorithmic decisions, autonomous machines).

To support data users in understanding this difficult landscape, ethical guidelines have been generated, and professional codes of conduct are being discussed among different communities of practice (http://aoir.org/reports/ethics2.pdf). Simultaneously, efforts to embed ethical thinking in the engineering and innovation community (e.g. value sensitive design (http://www.vsdesign.org/) and the responsible research and innovation frameworks (https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation)) are also being promoted to ensure technologies that are designed to anticipate consequences, mitigate risks and encourage "privacy by design". Privacy by design is an essential principle to establish privacy-aware computing environments. In this context, "consent" by a data subject to the processing of health-related data plays a key role. When applying Data Science, it will not be uncommon to process thousands or millions of health data points originating from data subjects. However, this processing must thus similarly respect thousands or millions of specific consent agreements to the processing of each subject's data. The need to automate such a verification process becomes obvious, and there are ongoing efforts (https://genomicsandhealth.org/working-groups/our-work/automatable-discovery-and-access) to represent consent data types in computer-readable format allowing for the automated discovery of accessible data across networked environments. In line with above, there have been also refined approaches enabling joint analysis of data without the need to share it, which are based on privacy-preserving data analytics techniques. Processing medical data brings major privacy challenges

[1] In the EU context, it has been pointed out that, even though the argument for free data flow and privacy are both strong, the latter prevails and the "solution must respect the rights of the individual to data protection, as laid down in the EU Charter, which also specifies that such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law" (EAPM 2013: 38).

in terms of who can process data and for what purpose. In particular, for joint analysis on data from different providers (e.g. hospitals), there is typically no single place in which the data can be collected and processed. Anonymization may require removing so much information from datasets that the quality of the analysis severely degrades. With privacy-preserving data analytics, on the other hand, different providers can contribute non-anonymized sensitive inputs to an analysis without the need to collect the data in one place. Smart use of encryption guarantees that no sensitive information leaves the provider—only the (non-sensitive) aggregated result of the analysis is shared.

5 Technology Landscape

This section provides a technology landscape on the application of Data Science to healthcare, in terms of (1) the technical challenges; (2) the various enabling platforms, services and infrastructures; and (3) data analytics methods, along with several success stories.

5.1 Technical Challenges

In the following, technical challenges and opportunities are discussed regarding the application of Data Science in healthcare.

5.1.1 Data Quality

There is the need to have reliable and reproducible results particularly in medical and pharmaceutical research where data gathering is extremely expensive. Data provenance provides an understanding of the source of the data—how it was collected, under which conditions but also how it was processed and transformed before being stored. This is important not only for reproducibility of analysis and experiments but also for understanding the reliability of the data that can affect outcomes in clinical and pharmacological research. As the complexity of operations grows, with new analysis methods being developed quite rapidly, it becomes key to record and understand the origin of data which in turn can significantly influence the conclusion from the analysis.

5.1.2 Data Quantity

The health sector is a knowledge-intensive industry depending on data and analytics to improve therapies and practices. There has been tremendous growth in the

range of information being collected, including clinical, genetic, behavioural, environmental, financial and operational data [47]. Healthcare data is growing at staggering rates that have not been seen in the past. There is a need to deal with this large volume and velocity of data to derive valuable insights to improve healthcare quality and efficiency. Organizations today are gathering a large volume of data from both proprietary data sources and public sources such as social media and open data. Through better analysis of these big data datasets, there is a significant potential to better understand stakeholder (e.g. patient, clinician) needs, optimize existing products and services as well as develop new value propositions. The breakthrough technologies, such as deep learning, require large quantities of data for training purposes. This data needs to come with annotations (ground truth). It is still very challenging in healthcare to arrange large quantity of representative data with high-quality annotations.

5.1.3 Multimodal Data

In healthcare, different types of information are available from different sources such as electronic healthcare records; patient summaries; genomic and pharmaceutical data; clinical test results; imaging (e.g. X-ray, MRI, etc.); insurance claims; vital signs from, e.g. telemedicine; mobile apps; home monitoring; ongoing clinical trials; real-time sensors; and information on wellbeing, behaviour and socioeconomic indicators. This data can be both structured and unstructured. The fusion of healthcare data from multiple sources could take advantage of existing synergies between data to improve clinical decisions and to reveal entirely new approaches to treat diseases [42]. For instance, the fusion of different health data sources could make the study and correlation of different phenotypes (e.g. observed expression of diseases or risk factors) possible that have proved difficult to accurately characterize from a genomic point of view only and thus enable the development of automatic diagnostic tools and personalized medicine. The combination and analysis of multimodal data poses several technical challenges related to interoperability, machine learning and mining.

Integration of multiple data sources is only possible if there are on the one hand de jure or de facto standards and data integration tooling and on the other hand methods and tools for integrating structured and unstructured (textual, sound, image) data. An example for the interoperability and data integration limitations is the relation between national and international health data standards. For example, in Germany, the xDT family of standards (ftp://ftp.kbv.de/ita-update/Abrechnung/KBV_ITA_VGEX_Datensatzbeschreibung_KVDT.pdf) is widely used by physicians and healthcare administration. xDT is not yet mapped to FHIR (http://hl7.org/implement/standards/fhir/index.html), its international counterpart in the HL7 framework. Without such a mapping, a Data Science solution will not be able to integrate the data fields relevant for a given analytics task.

5.1.4 Data Access

Although there is a sense of great opportunities regarding the analysis of health data for improving healthcare, there are very important barriers that limit the access and sharing of health data among different institutions (see the previous section on "Privacy, Ethics and Security") and countries. Political concerns, ethics and emotional aspects have a significant weight in this area. Privacy concerns form a very important aspect that needs to be overcome as well. There is a high degree of fragmentation in the health sector: collected data is not shared among institutions, even not within departments. This leads to the existence and spread of different isolated data silos that are not fully exploited. Insights cannot be derived from datasets that are disconnected. Top-down Data Science initiatives have not made much progress so far, and then several efforts are now focusing on a bottom-up approach. Changing the perspective to be patient-oriented gives patients more control over their data. Patients should thus be able to access their own data and decide whom to share it with and for what purpose. Examples are the social network PatientsLikeMe, which not only allows patients to interact and learn from other people with the same conditions but also provides an evidence base of personal data for analysis and a platform for linking patients with clinical trials.

5.1.5 Patient-Generated Data

Patient-generated health data (PGHD [16]) is defined as "health-related data including health history, symptoms, biometric data, treatment history, lifestyle choices which is created, recorded, gathered, inferred by, or from patients/caregivers to help address a health concern" (http://jop.ascopubs.org/content/early/2015/04/07/JOP.2015.003715.full#ref-3). This is differentiated from data generated during clinical care, because patients (not providers) are the ones responsible for capturing this data and also have the control over how this data are shared.

The proliferation of more affordable wearable devices, sensors and technologies such as patient portals to capture and transmit PGHD provides an unparalleled opportunity for long-term, persistent monitoring of the daily activities and responses of chronically ill patients. This engages patients as partners in their care allowing for advancements towards a true learning-based healthcare system for management of chronic diseases.

PGHD can help closing gaps in information and can offer healthcare providers a way to monitor a patient's health status and compliance to a therapy in between medical visits. It allows a way to gather information on a continuous basis rather than at a single point in time. Moreover, PGHD can provide the foundation for real-time care management programmes tailored to a single patient and their conditions. It can also aid in the management of chronic and acute conditions such as cardiac arrhythmias, congestive heart failure and diabetes. By providing relevant information about a patient's condition and health status, PGHD technologies can

encourage healthy behaviours and increase the success of preventive health and wellness programmes.

One of the largest concerns facing PGHD is in regard to data quality and provenance—i.e. the process of tracing and recording the quality and source of the data as it enters the system and moves across databases.

5.1.6 Usability/Deployment Methodology

Data Science holds tremendous promises for improving healthcare. But how should an organization get started with handling, organizing and analysing big data? Capitalizing on its opportunities requires an end-to-end strategy in which IT departments or groups are the technical enablers; but key executives, business groups and other stakeholders help setting objectives, identify critical success factors and make relevant decisions. Together these groups should consider existing problems that have been difficult to address as well as problems that have never been addressed before since data sources were unavailable or data was too unstructured to utilize. IT groups must solicit information from peers and vendors to identify the best software and hardware solutions for analysing big data in a healthcare context. Defining and developing use cases will help organizations focusing on the right solutions and creating the best strategies. As part of this process, IT groups should:

- map out data flows,
- decide what data to include and what to leave out,
- determine how different pieces of information relate to one another,
- identify the rules that apply to data,
- consider which use cases require real-time results and which do not, and
- define the analytical queries and algorithms required to generate the desired outputs.

They should define the presentation and analytic application layers, establish a data lake or warehousing environment and, if applicable, implement private- or public-based cloud data management. Some questions that should be asked are:

- What are the data requirements on collecting, cleansing and aggregating data?
- What data governance policies need to be in place for classifying data and meeting regulatory requirements?
- What infrastructure is needed to ensure scalability, low latency and performance?
- How will data be presented to business and clinical users in an easy-to-understand and easily accessible way?

5.2 Platforms, Services and Infrastructures

5.2.1 High-Performance Computers and Exascale Computing

There will be use cases, e.g. precision medicine, where the promises brought by Data Science will only be fulfilled through dramatic improvements in computational performance and capacity, along with advances in software, tools and algorithms. Exascale computers (HPCs)—machines that perform one billion calculations per second and are over 100 times more powerful than today's fastest systems—will be needed to analyse vast stores of clinical and genomic data. The use cases that will benefit the most from HPC—Data Science integration—are:

- **Precision medicine.** The new technology driving precision medicine is the area of omics. Omics data of a patient (genomics, metabolomics, proteomics, etc.) in combination with historical data about diseases and outcomes of different treatments allow making decisions whether a certain treatment would be beneficial for a patient, avoiding potential harming and the use of inefficient therapies. In life-threatening situations, these decisions need to be made in real time. Due to vast amount of data that needs to be analysed, the domain of precision medicine will benefit from using the HPC infrastructure and can help saving lives in an emergency department (ED).
- **Deep learning.** Deep learning algorithms have already shown a breakthrough performance in the medical domain. The advantage of deep learning algorithms is the capability that they can analyse very complex data, such as medical images, videos, text and other unstructured data. Deep learning algorithms will benefit from HPC infrastructure in cases when a large amount of data needs to be used for training of deep neural networks in order to provide relevant inputs to medical specialists as quickly as possible. One of the main areas where deep learning showed a tremendous potential is in the area of radiology. Deep learning algorithms can help in improving workflows within a hospital related to the diagnosis and treatments of the patients in the radiology department. This allows clinicians making quick decisions that would secure right and timely treatments of the patients.

5.2.2 Infrastructure

To manage and exploit this new flood of data, it is necessary to offer new infrastructures able to address the big data dimensions (i.e. volume, variety, veracity, velocity). In this respect, well-designed, solid and reliable infrastructures, which are not limited only to the IaaS level, provide the foundation on top of which all the other platforms and services can be provided. Advances offered by virtualization and cloud computing are today facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data [51] but will need to be more expansive to cope with the expected impact of future (healthcare)

data. The current cloud infrastructures are potentially ready to welcome the big data tsunami, and some technologies (e.g. Hadoop, Spark, MongoDB, Cassandra, etc.) are already going in this direction. Even if some requirements are satisfied, many issues still remain. Many applications and platforms, although used as services (SaaS/PaaS) directly from the cloud infrastructure, have not been designed to be dynamically scalable, to enable distributed computation, to work with nontraditional databases or to interoperate with infrastructures. For this reason, for (existing) cloud infrastructures, it will also be necessary to massively invest in solutions designed to offer dynamic scalability, infrastructure interoperability and massive parallel computing in order to effectively enable reliable execution of, for example, machine learning algorithms, pattern recognition of images, languages, media, artificial intelligence techniques, semantic interoperability and 3D visualization and other services. Furthermore, healthcare poses specific requirements on Data Science infrastructures (e.g. regulatory compliance, reliability, etc.).

Still there are several platforms and infrastructure in use in the healthcare sector. As an example, the Philips HealthSuite (http://www.usa.philips.com/healthcare/innovation/about-health-suite) [54] provides a cloud-based infrastructure for connected healthcare. With this platform, clinical and other data (from medical systems and devices) can be collected, combined and analysed. It enables care to become more personalized and efficient. Care providers and individuals are empowered to access (individual or aggregated) data on personal health, patient conditions and entire populations. Data from both the hospital and home are analysed with proprietary algorithms to identify health patterns and trends. This will lead to improved (clinical) decisions.

The importance of cloud computing was recently highlighted by the European Commission through its European Cloud Initiative (http://europa.eu/rapid/press-release_IP-16-1408_en.htm). They proposed a European Open Science Cloud; a trusted, open environment for the scientific community for storing, sharing and reusing scientific data and results; and a European data infrastructure targeting the build-up of the European supercomputing capacity. Data Science for the healthcare community must become an active partner supporting this initiative to ensure it accounts for its needs and that it serves the entire spectrum of professionals working in the field. In the following sections, further functionalities and features that the Data Science infrastructures should offer are described.

5.2.3 Data Integration

Data is being generated by different sources and comes in a variety of formats including unstructured data. All of this data needs to be integrated or ingested into big data repositories or data warehouses. This involves at least three steps, namely, extract, transform and load (ETL). With the ETL processes that have to be tailored for medical data have to identify and overcome structural, syntactic and semantic heterogeneity across the different data sources. The syntactic heterogeneity appears in the form of different data access interfaces, which were mentioned above, and

needs to be wrapped and mediated. Structural heterogeneity refers to different data models and different data schema models that require integration on schema level. Finally, the process of integration can result in duplication of data that requires consolidation.

The process of data integration can be further enhanced with information extraction, machine learning and Semantic Web technologies that enable context-based information interpretation. Information extraction will be a means to obtain data from additional sources for enrichment, which improves the accuracy of data integration routines, such as deduplication and data alignment. Applying an active learning approach ensures that the deployment of automatic data integration routines will meet a required level of data quality. Finally, the Semantic Web technology can be used to generate graph-based knowledge bases and ontologies to represent important concepts and mappings in the data. The use of standardized ontologies will facilitate collaboration, sharing, modelling and reuse across applications.

5.2.4 Interoperability Standards

In a data-driven healthcare environment, interoperability and standardization are key to deploy the full potential of data. However, there are still standardization problems in the healthcare sector since data is often fragmented or generated in IT systems with incompatible formats [56]. Research, clinical activities, hospital services, education and administrative services are organized in silos, and, in many organizations, each silo maintains its own separate organizational (and sometimes duplicated) data and information infrastructure. This poses barriers to combine and analyse data from different sources so as to identify insights and facilitate diagnosis. The lack of cross-border coordination and technology integration calls for standards to facilitate interoperability among the components of the Data Science value chain. As such, the creation of open, interoperable, patient-centred environments that promote rapid innovation and broad dissemination of advances is necessary as well as the promotion of open standards.

A large amount of terminological knowledge sources has been created in the realm of healthcare, e.g. the SNOMED clinical terms, the series of ICD classifications (ICD-9, ICD-10, etc.) or the Medical Subject Headings (MeSH) metathesaurus which is part of the Unified Medical Language System (UMLS (https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus)). Within SNOMED-CT, there are mappings between terms and also across languages. Since these knowledge sources are used in healthcare frameworks like HL7, a data analytics system must be able to process them in (cross-lingual) indexing and retrieval scenarios. Hence, there is a need for:

- tooling that allows processing and integrating these knowledge sources in a given healthcare framework and that can be deployed in different Data Science healthcare workflows;

- and guidelines and best practices that inform providers and users of healthcare data on adequate processes and workflows, for handling knowledge systems in healthcare.

In addition to terminology, there are several other areas with interoperability challenges (http://www.lider-project.eu/sites/default/files/D3.2.2-Phase-II.pdf). For laboratory analytical processes, the Allotrope Foundation[2] is developing a common vocabulary and file format to support exchange of laboratory data. For the reuse of patient data, not only technical challenges but also regulatory and legal frameworks make data sharing extremely difficult. A general concern is the language barrier. Many knowledge systems like ICD or SNOMED-CT have a restricted set of multilingual labels. Reusing the knowledge systems in another language or health system comes with high costs.

In the realm of PGHD, the lack of industry-wide standards is a growing concern within the information technology community. Although many device companies are using standards profiled by Continua Health Alliance or the consolidated care document (CCD) standard (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258) that enables connectivity between sources, many devices (such as the popular "Fitbit" device) still use proprietary architectures and formats making it more difficult for interoperability given that patients may have multiple devices.

Integrating outside data sources (like PGHD) into the EHR is difficult because there are no industry standards for this activity and EHRs are often designed to be proprietary. This can have a significant impact on both project time and cost. Industry standards organizations such as HL7 are actively working on these issues and especially on standard methods for capturing PGHD, recording PGHD and making PGHD interoperable within the current framework of structured documents. Common health IT standards and terminologies should be leveraged where possible—e.g. LOINC for lab results and RxNorm for medication terminologies—however, it is likely that, due to the demands and needs of the various stakeholders involved (patients, providers, EHR vendors, application developers, etc.), new standards will have to be developed. Since healthcare recommendations, standards and policies are constantly evolving, flexibility should be built into the new technology to allow for rapid response to change.

5.3 Data Analytics

Medical research has always been a data-driven science, with randomized clinical trials being a gold standard in many cases. However, due to recent advances in omics technologies, medical imaging, comprehensive electronic health records and

[2] http://www.allotrope.org/.

smart devices, medical research and clinical practice are quickly changing into data-driven fields. As such, the healthcare domain as a whole—doctors, patients, management, insurance and politics—can significantly profit from current advances in Data Science, and in particular from data analytics.

There are certain challenges and requirements to develop specialized methods and approaches for data analytics in healthcare. These include:

- **Multimodal data**:
  Optimally, in data analytics, there is a set of well-curated, standardized and structured data—for example, as sometimes found in electronic health records. However, a high percentage of health data is a variety of unstructured data. Much of it comes in forms of real-time sensor readings such as ECG measurements in intensive care, text data in clinical reports by doctors, medical literature in natural language, imaging data or omics data in personalized medicine. Furthermore, the use of external data such as lifestyle information, e.g. for disease management, or geospatial data and social media for epidemiology is becoming increasingly common. It is vital to gain knowledge from that information. The goal should be to obtain valuable information from such heterogeneous data through multimodal learning, make the insights from such combined information available to clinicians and incorporate knowledge into the clinical history of patients.
- **Complex background knowledge**:
  Medical data needs to describe very complex phenomena, from multi-level patient data on medical treatment and procedures, lifestyle and information to the vast amount of available medical knowledge in the literature, biobanks or trial repositories. Hence, medical data usually comes with complex metadata that needs to be taken into account in order to optimally analyse the data, draw conclusions, find appropriate hypotheses and support clinical decisions.
- **Explainable trustworthy models**:
  End users of analytical tools in medicine—such as doctors, clinical researchers and bioinformaticians—are highly qualified. They also have a high responsibility, from which follow high expectations on the quality of analytics tools before trusting them in the treatment of patients. Hence, an optimal analytical approach should, as much as possible, generate understandable patterns in order to allow for cross-checking results and enabling trust in the solutions. It should also enable expert-driven self-service analytics to allow the expert to control the analytics process.
- **Supporting complex decision**:
  The analysis of imaging data, pathology, intensive care monitoring and the treatment of multi-morbidities are examples of areas in which medical decisions have to be taken from noisy data, in complex situations, and with possibly missing information. Neither humans nor algorithms may be guaranteed to always deliver an optimal solution, yet they may be required to take important decisions or specify options in minimal time. Another area of medical decision support with potentially very high future impact is smart assistants for patients that make use of smartphones and new wearable devices and sensor technologies to help patients manage diseases and lead healthier lives.

- **Privacy**:
  Medical data is a highly sensitive information that is protected by strong legal safeguards at the European level. An adequate legal framework to enable the analysis of such data, and the development of adequate privacy-preserving analytical tools to implement this framework, is of high importance for the practical applicability and impact of data-driven medicine and healthcare.

Approaches to address data analytics under the aforementioned challenges are presented in the following.

5.3.1 Advanced Machine Learning and Reinforcement Learning

Many healthcare applications would significantly benefit from the processing and analysis of multimodal data—such as images, signals, video, 3D models, genomic sequences, reports, etc. Advanced machine learning systems [37] can be used to learn and relate information from multiple sources and identify hidden correlations not visible when considering only one source of data. For instance, combining features from images (e.g. CT scans, radiographs) and text (e.g. clinical reports) can significantly improve the performance of solutions.

The fusion of different health data sources could also enable the study of phenotypes (e.g. diseases or risk factors) that have proven difficult to characterize from a genomic point of view only. This will enable the development of automatic diagnostic tools and personalized medicine. This technology will be key to leverage the full potential of the varied sources of big data.

Another aspect is the analysis of lifestyle data collected from apps on smartphones and from which may include information about risk factors for diseases and disease management such as specialized hardware, activity information, GPS tracks and mood tracking, which can otherwise not be reliably collected. This information can be used within (learning) recommender systems that help monitoring patients, raise alarms or give advice for the better handling of a disease.

Reinforcement learning is a new very promising advanced machine learning method with a paradigm of learning by trial-and-error, solely from rewards or punishments. It was successfully applied in breakthrough innovation, such as AlphaGo system of DeepMind that won the Go game against the best human player. It can also be applied in the healthcare domain, for example, to dynamically optimize workflows.

5.3.2 Deep Learning

Deep learning [13] typically refers to a set of machine learning algorithms based on learning data representations (capturing highly nonlinear relationships of low-level unstructured input data to form high-level concepts). Deep learning approaches [3] made a real breakthrough in the performance of several tasks with which traditional

machine learning methods were struggling such as speech recognition, machine translation, computer vision (object recognition), etc. For example, they are nowadays a preferred method in medical image analysis, allowing medical specialists, who depend on insights from medical images, e.g. radiologists or pathologists, to quickly analyse these images.

Deep hierarchical models are artificial neural networks (ANN) with deep structures and related approaches, such as deep restricted Boltzmann machines, deep belief networks and deep convolutional neural networks. The current success of deep learning methods is enabled by advances in algorithms and high-performance computing technology, which allow analysing the large datasets that have now become available.

5.3.3 Real-Time Analytics

Certain time-critical healthcare applications need actions to be taken right at the point when a particular event is detected (e.g. alarms in the ICU). Multiple streams of heterogeneous data offer the possibility to extract insights in real time. Several relevant, interconnected approaches exist:

- Real-time analytics refers to analytics techniques, which can analyse and create insights from all available data and resources in real time as they come into a system.
- Data stream mining refers to the ability to analyse and process streaming data in the present (or as it arrives), rather than storing the data and retrieving it at some point in the future.
- Complex event detection refers to the discovery and management of patterns over multiple data streams, where patterns are high-level, semantically rich and made ultimately understandable to the user.

5.3.4 Clinical Reasoning

There is the need to improve clinical decisions by incorporating information derived from various forms of human input (e.g. free text, voice input, medical records, medical ontologies, etc.) and where semantics can be used to facilitate this [20]. Scientific insights from cognitive science, neuroanatomy and neurophysiology have resulted in the generation of mathematical models that can simulate large multilayer and nonrandom networks of components for data processing and inferencing to accomplish complex tasks such as automated reasoning and decision-making. Clinical reasoning leverages various techniques including distributed information representation, machine learning, natural language processing (NLP), semantic reasoning, statistical inferencing, fuzzy logic, image processing, signal processing and the synaptic-type communications in biological neurons. Artificial neural networks which are, essentially, models of unsupervised learning in a cognitive system with

hidden layers representing "weighted" connections and fault tolerance similar to thought processes in animals and humans are critical to cognitive computing [18].

5.3.5 End User-Driven Data Analytics

End user-driven data analytics—which is also becoming more and more prominent under the name Citizen Data Science (http://www.gartner.com/newsroom/id/3114217)—enables the average user to make use of modern analytical solutions. The user in this case may be a patient or a very experienced domain expert—a doctor, hospital management staff, biological researcher, etc.—but without an in-depth knowledge of statistics, data processing and methods and tools. Approaches for end user-driven analytics include visual and interactive analytics. More and more question-answering approaches that allow a party to phrase more complex natural language questions are reaching maturity. The availability of such smart, easy-to-use tools enables professionals to make use of data-driven decision-making on all levels. A particular case of end user-driven analytics may be found in the phenomenon of the "quantified self", where patients collect much data about themselves and analyse it to find insights about their health status or disease.

5.3.6 Natural Language Processing and Text Analytics

From the perspectives of data content processing and data mining, textual data belongs to so-called unstructured data just as images or videos because of the complexities of their internal structures. Technologies such as information retrieval and text analytics have been created for facilitating easy access to this wealth of textual information. Text analytics is a broad term referring to technologies and methods in computational linguistics and computer science for the automatic detection and analysis of relevant information in unstructured textual content (free text). Often machine learning and statistical methods are employed for text analytics tasks. In the literature, text analytics is also regarded as a synonym of (1) text mining or (2) information and knowledge discovery from text. Major subtasks are (1) linguistic analysis, (2) named entity recognition, (3) coreference resolution, (4) relation extraction and (5) opinion and sentiment analysis [21, 53]. In the context of language processing and text analytics, there are several tools that have been widely used for the extraction of knowledge from biomedical and clinical natural text such as MetaMap [2], Apache cTAKES [57] or NCBO Annotator [36], among others. The number of approaches in this area is really vast [25, 45, 52], and they are different in specific domains (phenotype extraction, gene extraction, protein interactions, etc.). However, most existing models, tools and corpora focus on English data only, which makes the processing of non-English biomedical or clinical text more difficult. Even though various non-English datasets exist (e.g. French [44], German [55], Spanish [12] or Swedish [60]) which are required to train extraction models, datasets are often not publicly available due to legal regulations.

There is a strong need to improve clinical decisions by incorporating semantics derived from various forms of human input (e.g. free text, medical records, literature). Vast amount of information is currently held in medical records in the form of free text. Thus, text analytics is important to unravel the insights within the textual data. Particularly in healthcare, but in almost all other industries, records (digital or not) are still kept as free text. There is plethora of applications in the clinical setting where practitioners produce and rely on free text for reporting diagnosis and operations. Of particular importance is the mining of medical literature [34], which enables the use of vast amounts of medical knowledge more efficiently. Examples include literature recommender systems and also the detection of new medical knowledge from literature, e.g. for drug repositioning [17].

Given the large amount of biomedical knowledge recorded in textual form, full papers, abstracts and online content, there is the need for techniques that can identify, extract, manage and integrate this knowledge. In parallel, text analytics tools have been adapted and further developed for extracting relevant concepts and relations among concepts from clinical data such as patient records or reports written by doctors. The information extraction technology plays a central role for text mining and text analytics. Even though there has been significant breakthrough in natural language processing with the introduction of advanced machine learning technologies (in particular, recently, deep learning), these technologies need to be further developed to meet the challenges of large volumes and velocities.

5.3.7 Knowledge-Based Approaches

With the advent of the Semantic Web, description logics have become one of the most prominent paradigms for knowledge representation and reasoning. In medicine, the use of knowledge bases constructed from sophisticated ontologies has proven to be an effective way to express complex medical knowledge and support the structuring, quality management and integration of medical data. Also the mining of other complex data types, such as graphs [19] and other relational structures, is motivated by various applications in biological networks such as pathways or in secondary structures of macromolecules such as RNA and DNA.

These and many other occurrences of data are arising and growing. Learning from this type of complex data can hence yield more concise, semantically rich, descriptive patterns in the data which better reflect its intrinsic properties. In this way, discovered patterns promise more clinical relevance.

A complex analysis and multidisciplinary approach to knowledge is essential to understand the impact of various factors on healthcare systems. The challenges for understanding and addressing the issues concerning the healthcare world are the use of big data, non-conformance to standards and heterogeneous sources (in heterogeneous documents and formats), which need an immediate attention towards multidisciplinary complex data analytics on top of rich semantic data models. Ontology-driven systems result indeed in the effective implementation of healthcare strategies for the policymakers. The creation of semantic knowledge bases for

healthcare has an extremely high potential and practical impact. They facilitate data integration from multiple heterogeneous sources, enable the development of information filtering systems and support knowledge discovery tasks. In particular, in the last years the linked open data (LOD) initiative reached significant adoption and is considered the reference practice for sharing and publishing structured data on the web [7, 8]. LOD offers the possibility of using data across different domains for purposes like statistics, analysis, maps and publications. By linking this knowledge, interrelations and associations can be inferred, and new conclusions arise.

Healthcare data is generated in various sources in diverse formats using different terminologies. Due to the heterogeneous formats and lack of common vocabulary, the accessibility of the healthcare data is very minimal for health data analytics and decision support systems. Vocabulary standards are used to describe clinical problems and procedures, medications and allergies [11]. Important examples are, just to name a few, the Logical Observation Identifiers Names and Codes (LOINC), International Classification of Diseases (ICD9 and ICD10), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), Current Procedural Terminology, 4th Edition (CPT 4), Anatomic Therapeutic Chemical (ATC) Classification of Drugs, Gene Ontology (GO), RxNorm, General Equivalence Mappings (GEMs), OBO Foundry (http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html) and others.

Since healthcare systems are characterized by a large amount of data, heterogeneous in nature and with different quality and security requirements, research on data reengineering, linking, formalization and consumption is of primary interest. The heterogeneity problem has to be tackled at different levels. On the one hand, syntactic interoperability is needed to unify the format of knowledge sources enabling, e.g. distributed query [10]. Syntactic interoperability can be achieved by conforming to universal knowledge representation languages and by adopting standard practices. The widely adopted RDF, OWL and LOD approaches support syntactic interoperability. On the other hand, semantic interoperability is also needed. Semantic interoperability can be achieved by adopting a uniform data representation and formalizing all concepts into a holistic data model (conceptual interoperability). RDF and OWL assist in achieving the former goal. However, conceptual interoperability is domain specific and cannot be achieved only by the adoption of standard tools and practices but also through interlinking with existing healthcare knowledge bases by means of domain experts and semiautomated solutions. The large, heterogeneous data sources in the healthcare world make the problem even harder, as different semantic perspectives must be addressed in order to cope with knowledge source conceptualizations.

Once interoperability at both syntactic and conceptual levels is obtained, it is possible to intercross data and exploit them more in depth, providing application developers the opportunity to easily design their services and applications. Semantic interoperability at the domain level allows making sense of distributed data and enabling their automatic interpretation. In this way, the issue of resolving semantic interoperability among different data sources is moved from the application level to the data model level. Developers are then relieved from the burden of reconciling,

uniforming and linking data at a conceptual level and are able to build their solutions in a more intuitive and efficient way. The published data sources are made discoverable and become accessible via queries and/or public facilities and integrated into higher-level services.

Presently several international organizations and agencies across the world, e.g. the World Health Organization (WHO), make use of semantic knowledge bases in healthcare systems to:

- Improve accuracy of diagnoses by providing real-time correlations of symptoms, test results and individual medical histories;
- Help to build more powerful and more interoperable information systems in healthcare;
- Support the need of the healthcare process to transmit, reuse and share patient data;
- Provide semantic-based criteria to support different statistical aggregations for different purposes;
- Bring healthcare systems to support the integration of knowledge and data.

Putting knowledge management systems in place for healthcare can facilitate the flow of information and result in better, more-informed decisions.

5.3.8 High-Performance Genome Analysis

Currently, clinical applications of next-generation sequencing primarily focus on exome sequencing (about 2–3% of the genome) and more targeted assays, such as diagnostic panels of cancer genes. On the other hand, whole-genome analysis (WGS) delivers a complete and integrated view of the genome of a patient and as such can advantageously replace complex series of older ad hoc targeted genetic tests in areas such as cancer genomics, rare genetic diseases, preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS) or non-invasive prenatal testing (NIPT). WGS is quickly becoming economically competitive and promises that a correct diagnosis can be reached more often and more quickly. With thousands of patients in need of WGS at any large hospital each year, the computing and storage needs for clinical applications of WGS are set to explode.

Most genome analysis software was created incrementally during the last decade targeting single computer systems. Because of this legacy, the software is not optimized for throughput, concurrency and parallelism which is needed to achieve good performance and efficient usage of server-based systems. Running this software—as is—on a shared backend compute cluster will result in slow runtimes and too costly usage of the compute infrastructure.

WGS software pipelines consist of a number of software applications where each runs part of the computation. These applications tend to be written by different teams and in different programming languages. The most common interface between these applications in use today is raw data files. This results in severe performance penalties. Recently, a number of methods have been introduced to accelerate read mapping and variant calling through the use of high-performance and

distributed computing techniques (e.g. HugeSeq [40], Halvade [15], ADAM [46] or elPrep [33]). There is still a long road ahead to optimize the complete WGS software stacks, including the analysis tools, and get them fully adopted in practice.

5.3.9 Understanding and Reliability in Analytics

Often in medical decision-making, important—often literally life or death—decisions must be taken under time pressure and in complex and unclear situations with potentially severe consequences of errors if the right decision is not made. Even while recognizing that a data-driven approach may never be 100% correct, and even while considering that neither are human doctors always right, very high standards are required for data analytics in medical applications. Measuring and managing the performance (e.g. accuracy of data-driven systems) are therefore of utmost importance. Not only this is a basic ethical requirement, but the uptake of novel smart solutions into clinical practice is often hindered by unaddressed questions of liability and safety.

Key features of an analytical solution that inspire trust in its practical use are understanding—in particular enabling the human doctor or researcher to be aware of its advantages and limits and reliability—in particular for complex learning systems that evolve over time from a stream of new input data, guaranteeing reliability has been recognized as a major challenge [59].

As a consequence, understanding and reliability should be particularly addressed as a basic requirement in all applications of data analytics in medicine and healthcare.

5.4 Example Success Stories

- **Precision medicine initiative (https://www.whitehouse.gov/precision-medicine) launched by President Obama:**
  By taking into account individual differences in people's genes, environments and lifestyles, treatments can be tailored to the individual instead of applying a one-size-fits-all approach designed for the average patient. Six personal stories (https://www.whitehouse.gov/blog/2015/01/29/precision-medicine-already-working-cure-americans-these-are-their-stories) describe how precision medicine has led to a successful outcome with a personalized treatment.
- **European Medical Information Framework (EMIF) [23]:**
  An IMI project with a common platform for the reuse of clinical information is funded with 60 million EUR. It includes clinical information of about 50 million patients around Europe.
- **Open PHACTS Discovery Platform [48]:**
  Also funded by IMI, the platform integrates and links information from the most important drug and compound databases.

- **Integration of clinical research networks conforming Data Science repositories:**
  The value of integrating clinical research networks is widely recognized by researchers and funding agencies, since connecting networks means clinical research can be conducted more effectively, conforming communities with shared operational knowledge and data. Examples are the Li Ka Shing Centre for Health Information and Discovery of the University of Oxford, recently supported by a £90 M initiative in Data Science and drug discovery [49], or the NIH Big Data to Knowledge (BD2K) initiative [6] enabling biomedical scientists to capitalize on the data being generated by the research communities.
- **Philips HealthSuite digital platform:**
  HealthSuite offers both a native cloud-based infrastructure and the core services needed to develop and run a new generation of connected healthcare applications. Unlike other digital platforms, HealthSuite is built on purpose for the complex challenges of healthcare, featuring deep clinical databases, patient privacy, industry standards and protocols and personal and population data visualizations. This empowers healthcare providers to efficiently impact patient care.

6 Conclusions and Recommendations

This chapter has shown that there is a lot of potential in delivering more targeted, wide-reaching, and cost-efficient healthcare by exploiting Data Science and AI technologies. However, it has also been shown that the healthcare domain has some very specific characteristics and challenges that require a targeted effort and research in order to realize the full potential:

- *Data access, availability and quality*: There is a huge amount of existing data distributed in several repositories and new data generated daily by billions of connected devices or self-generated by people. It is necessary to find more appropriate and effective ways to leverage these data in line with privacy and ethical principles, to access it, to understand the purposes for its use and quality in order to improve and optimize care processes, disease diagnosis, personalized care and in general the healthcare system. However, in the healthcare sector, data is often fragmented or generated in different systems with incompatible formats. Therefore, interoperability and standardization are key to deploy the full potential of data.
- *Patients and healthcare professionals profiting from Data Science*: There is the need to develop approaches that allow for humans and machines to cooperate more closely on exploiting Data Science for a better health. This includes guarantees on the trustworthiness of information, a focus on generating actionable advice and improving the interactivity and understandability of data processing

and analytics. The requirements of different target groups—researchers, doctors and caregivers or patients and general population—may demand different focus.

- *Multimodal data analytics*: There is the need for technologies, which can handle, analyse and exploit the set of very diverse, interlinked and complex data that already exists in the healthcare universe to improve healthcare quality and decrease healthcare costs.
- *Healthcare knowledge*: Next to the big and heterogeneous healthcare datasets, there is already a big amount of medical and healthcare knowledge. This knowledge exists in books and research papers but also in the heads of healthcare professionals. In fields such as epidemiology or wearable sensors, also completely different knowledge on the real world, organizations and how people live their lives is very valuable to understand patients and the healthcare system in general. New approaches are needed that bring together data-driven and knowledge-based approaches, such that knowledge can be used to make better sense of data, and data can be used to generate more knowledge.
- *Ethics and privacy in Data Science*: Further practical approaches are needed to adequately balance the benefit and threats of more and more detailed and sensitive data being available. With respect to an increasing amount of complexity and automation in clinical data processing and decision support, and in particular in the light of the move towards personal health assistant on smartphones, a targeted focus on the ethical problems connected with these new technologies seems advisable.
- *Increasing focus on primary and secondary prevention*: Currently, 97% of healthcare budgets are spent on treating patients both with acute and chronic conditions (https://euobserver.com/chronic-diseases/125922). Only 3% is spent on prevention, with chronic diseases being among the most preventable illnesses. Considering the economic impact of chronic diseases on the productivity of the EU workforce, an increased focus on primary and secondary prevention is clearly needed, and Data Science and AI are geared to help here.
- *Policies and technologies towards digital single market strategy*: Dealing with different health data protection regimes across EU Member States creates difficulties in accessing and sharing health data at EU level. The implementation of the GDPR is an opportunity to look for alignment. Finally, innovative approaches to healthcare, such as value-based healthcare, should be supported by policy to drive the transformation of the healthcare sector. Developing policies and technologies will contribute towards enabling the digital single market strategy.

To prove the impact of Data Science and AI technologies on the healthcare sector, it is essential to apply these recommendations in large-scale pilots. The pilots are meant to serve as the best practice examples. Their objective is to demonstrate how the health sector can be transformed with the aim to increase its quality, decrease costs and improve accessibility. This can be done by putting Data Science technologies at their core with the goal that their results can be scaled up and adopted by the whole healthcare sector.

References

1. A third of European hospitals report operating losses, according to Accenture nine-country study. https://newsroom.accenture.com/industries/health-public-service/a-third-of-european-hospitals-report-operating-losses-according-to-accenture-nine-country-study.htm
2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association, Bethesda (2001)
3. Atzeni, M., Recupero, D.R.: Deep learning and sentiment analysis for human-robot interaction. In: The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, June 3–7, 2018. Revised Selected Papers, pp. 14–18 (2018)
4. Auffray, C., et al.: Making sense of big data in health research: towards an eu action plan. Genome Med. **8**, 71 (2016)
5. Baro, E., Degoul, S., Beuscart, R., Chazard, E.: Toward a literature-driven definition of big data in healthcare. BioMed. Res. Int. **2015**, 639021 (2015)
6. Bd2k Mission Statement (2012). http://datascience.nih.gov/bd2k/about
7. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D.: Exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop, SWUI 2006, Athens (2006)
8. Berners-Lee, T., Bizer, C., Heath, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**, 1–22 (2009)
9. Big Data and Analytics for Infectious Disease Research, Operations, and Policy: Proceedings of a Workshop (2016). https://www.nap.edu/read/23654/chapter/1
10. Bizer, C., Heath, T.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web edition, vol. 344. Morgan & Claypool Publishers, San Rafael (2011)
11. Colin, P., Karthik, P.G., Preteek, J., Peter, Y., Kunal, V.: Multiple ontologies in healthcare information technology: motivations and recommendation for ontology mapping and alignment. In: Proceedings of International Conference on Biomedical Ontologies, New York, pp. 367–369 (2011)
12. Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of entities and relations in Spanish radiology reports. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, pp. 177–184. INCOMA Ltd, Moskva (2017)
13. Courville, A., Goodfellow, I., Bengio, Y.: Deep Learning (2016). http://www.deeplearningbook.org
14. Data silos: Healthcare's silent shame. http://www.forbes.com/sites/davidshaywitz/2015/03/24/data-silos-healthcares-silent-tragedy/#19b0f7f99394
15. Decap, D., Reumers, J., Herzeel, C., Costanza, P., Fostier, J.: Halvade: scalable sequence analysis with mapreduce. Bioinformatics **31**(15), 2482–2488 (2015)
16. Deering, M.J.: Issue brief: patient-generated health data and health it. The Office of the National Coordinator for Health Information Technology (2013)
17. Deftereos, S.N., Andronis, C., Friedla, E.J., Persidis, A., Persidis, A.: Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdiscip. Rev. Syst. Biol. Med. **3**(3), 323–334 (2011)
18. Dessì, D., Reforgiato Recupero, D., Fenu, G., Consoli, S.: Exploiting cognitive computing and frame semantic features for biomedical document clustering, vol. 1948, pp. 20–34 (2017). Cited By 4
19. Dessì, D., Cirrone, J., Recupero, D.R., Shasha, D.E.: Supernoder: a tool to discover over-represented modular structures in networks. BMC Bioinf. **19**(1), 318:1–318:12 (2018)
20. Dessì, D., Reforgiato Recupero, D., Fenu, G., Consoli, S.: A recommender system of medical reports leveraging cognitive computing and frame semantics. Intell. Syst. Ref. Libr. **149**, 7–30 (2019). Cited By 0

21. Dridi, A., Reforgiato Recupero, D.: Leveraging semantics for sentiment polarity detection in social media. Int. J. Mach. Learn. Cybern. (2017). https://doi.org/10.1007/s13042-017-0727-z
22. European Centre for Disease Prevention and Control. http://ecdc.europa.eu/en/healthtopics/Healthcare-associated_infections/Pages/index.aspx
23. European Medical Information Framework (EMIF). http://www.emif.eu
24. Garcia-Barbero, M., Gröne, O.: Trends in integrated care reflections on conceptual issues. World Health Organization, Copenhagen, EUR/02/5037864 (2002)
25. Hahn, U., Cohen, K.B., Garten, Y., Shah, N.H.: Mining the pharmacogenomics literature survey of the state of the art. Brief. Bioinform. **13**(4), 460–494 (2012)
26. Hai Data and Statistics, Centers for Disease Control and Prevention (2016). http://www.cdc.gov/HAI/surveillance/
27. Health at a glance 2015, OECD indicators. http://www.oecd-ilibrary.org/social-issues-migrationhealth/health-at-a-glance-2015/summary/english_47801564-en;jsessionid=fnol3e9ktakqk.x-oecd-live-03
28. Healthcare Breach Report, Bitglass Report (2016). Available at: http://pages.bitglass.com/rs/418-ZAL-815/images/BR_Healthcare_Breach_Report_2016.pdf
29. Healthcare data growth: an exponential problem. http://www.nextech.com/blog/healthcare-data-growth-an-exponential-problem
30. Health care systems: getting more value for money. http://www.oecd.org/eco/growth/46508904.pdf
31. Health and health systems. http://ec.europa.eu/europe2020/pdf/themes/05_health_and_health_systems.pdf?_sm_au_=iHVqq23HLDVwQ7DP
32. Healthy aging data and statistics. http://www.euro.who.int/en/health-topics/Life-stages/healthy-ageing/data-and-statistics
33. Herzeel, C., Costanza, P., Decap, D., Fostier, J., Reumers, J.: elPrep: high-performance preparation of sequence alignment/map files for variant calling. PLOS One **10**(7), e0132868 (2015). https://doi.org/10.1371/journal.pone.0132868
34. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: state-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer, Berlin (2014)
35. Investing in health. http://ec.europa.eu/health/strategy/docs/swd_investing_in_health_en.pdf
36. Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M.-A., Musen, M.: NCBO annotator: semantic annotation of biomedical data. In: International Semantic Web Conference, Poster and Demo session, vol. 110 (2009)
37. Khosla, A., Ngiam, J., et al.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA (2011)
38. Kissick, W.: Medicine's Dilemmas. Yale University Press, New Haven (1994)
39. Kou, S.C., Yang, S., Santillana, M.: Accurate estimation of influenza epidemics using google search data via argo PNAS (2015). http://www.pnas.org/content/112/47/14473
40. Lam, H.Y., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D., Snyder, M.: Detecting and annotating genetic variations using the hugeseq pipeline. Nat. Biotechnol. **30**(3), 226–229 (2012)
41. Luo, B., Sampathkumar, H., Chen, X.-W.: Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC Med. Inform. Decis. Mak. **14**, 91 (2014)
42. May, M.: Life science technologies: big biological impacts from big data. Science **344**(6189), 1298–1300 (2014)
43. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Trends in integrated care reflections on conceptual issues. Big data: the next frontier for innovation, competition, and productivity, McKinsey Global Institute Technical Report. Available at: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
44. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, September 8–11 (2015)

45. Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. Brief. Bioinform. **15**(2), 327–340 (2012)
46. Nothaft, F.: Scalable genome resequencing with Adam and Avocado. Master's thesis, EECS Department, University of California, Berkeley (2015)
47. OECD: Data-Driven Innovation: Big Data for Growth And Well-Being. OECD Publishing, Paris (2015)
48. Openphacts bringing together pharmacological data resources in an integrated, interoperable infrastructure. http://openphacts.org
49. Oxford, U.O. prime minister joins sir ka-shing li for launch of 90m initiative in big data and drug discovery at oxford university (2014). http://www.ox.ac.uk/media/news_releases_for_journalists/130305.htm
50. Personal health train architecture for analyzing distributed data repositories. http://www.dtls.nl/fair-data/personal-health-train/
51. Raghupathi, V., Raghupathi, W.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**, 3 (2014)
52. Rebholz-Schuhmann, D., Oellrich, A., Hoehndorf, R.: Text-mining solutions for biomedical research: enabling integrative biology. Nat. Rev. Genet. **13**(12), 829–839 (2012)
53. Recupero, D.R., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A.G.: Sentilo: frame-based sentiment analysis. Cogn. Comput. **7**(2), 211–225 (2015)
54. Rodriguez, M.L., Quelch, J.A.: Philips healthcare: marketing the healthsuite digital platform. Harvard Business School Case 515-052 (2015). https://hbr.org/product/Philips-Healthcare--Marke/an/515052-PDF-ENG (Revised September 2015)
55. Roller, R., Rethmeier, N., Thomas, P., Hübner, M., Uszkoreit, H., Staeck, O., Budde, K., Halleck, F., Schmidt, D.: Detecting Named Entities and Relations in German Clinical Reports, pp. 146–154. Springer, Cham (2018)
56. Roney, K.: If interoperability is the future of healthcare, what's the delay? Becker's Hospital Review (2012). Available at: https://www.beckershospitalreview.com/healthcare-information-technology/if-interoperability-is-the-future-of-healthcare-whats-the-delay.html
57. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inf. Assoc. **17**(5), 507–513 (2010)
58. Scott, R.D., II.: The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention. Stephen B. Thacker CDC Library Collection, document number cdc:11550. Available at: https://stacks.cdc.gov/view/cdc/11550
59. Sculley, D., et al.: Hidden technical debt in machine learning systems. In: Proceedings of Neural Information Processing Systems (NIPS) (2015)
60. Skeppstedt, M., Kvist, M., Nilsson, G.H., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J. Biomed. Inf. **49**, 148–158 (2014)
61. Tackling chronic disease in Europe strategies, interventions and challenges. http://www.euro.who.int/__data/assets/pdf_file/0008/96632/E93736.pdf
62. Teisberg, E.O., Porter, M.E.: Redefining Health Care: Creating Value-Based Competition on Results. Harvard Business Press, Boston (2006)
63. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016). http://www.nature.com/articles/sdata201618