Taylor & Francis
Taylor & Francis Group

Expert Review of Precision Medicine and Drug Development

Personalized medicine in drug development and clinical practice

ISSN: (Print) 2380-8993 (Online) Journal homepage: http://www.tandfonline.com/loi/tepm20

The path from big data to precision medicine

Bevan E Huang, Widya Mulyasasmita & Gunaretnam Rajagopal

To cite this article: Bevan E Huang, Widya Mulyasasmita & Gunaretnam Rajagopal (2016) The path from big data to precision medicine, Expert Review of Precision Medicine and Drug Development, 1:2, 129-143, DOI: 10.1080/23808993.2016.1157686

To link to this article: http://dx.doi.org/10.1080/23808993.2016.1157686

Accepted author version posted online: 29 Feb 2016.
Published online: 09 Mar 2016.

EXPERT REVIEW OF PRECISION MEDICINE AND DRUG DEVELOPMENT, 2016
VOL. 1, NO. 2, 129–143
http://dx.doi.org/10.1080/23808993.2016.1157686

Taylor & Francis
Taylor & Francis Group

REVIEW

The path from big data to precision medicine

Bevan E Huang[a], Widya Mulyasasmita[b] and Gunaretnam Rajagopal[c]

[a]Discovery Sciences, Janssen R&D, Menlo Park, CA, USA; [b]Johnson & Johnson Innovation, Menlo Park, CA, USA; [c]Discovery Sciences, Janssen R&D, Spring House, PA, USA

ABSTRACT
Precision medicine aims to combine comprehensive data collected over time about an individual's genetics, environment, and lifestyle, to advance disease understanding and interception, aid drug discovery, and ensure delivery of appropriate therapies. Considerable public and private resources have been deployed to harness the potential value of big data derived from electronic health records, 'omics technologies, imaging, and mobile health in advancing these goals. While both technical and sociopolitical challenges in implementation remain, we believe that consolidating these data into comprehensive and coherent bodies will aid in transforming healthcare. Overcoming these challenges will see the effective, efficient, and secure use of big data disrupt the practice of medicine. It will have significant implications for drug discovery and development as well as in the provisioning, utilization and economics of health care delivery going forward; ultimately, it will enhance the quality of care for the benefit of patients.

ARTICLE HISTORY
Received 3 January 2016
Accepted 19 February 2016
Published online
9 March 2016

KEYWORDS
Big data; EHR; mHealth; genomics; data integration; precision medicine; real world evidence; data analytics

The US FDA defines precision medicine as 'the tailoring of medical treatment to the individual characteristics, needs, and preferences of a patient during all stages of care, including prevention, diagnosis, treatment, and follow-up'. Big data, as it applies to precision medicine, is the generation and repository of large amounts of data from bio-specimens, health records, medical imaging and sensors, from which disease-specific factors, patterns, and associations can be computationally identified and used to customize medical treatments unique to the individual [1a].

Recently, big data and personalized or precision medicine have been heavily hyped, with 'data scientist' being touted as 'the sexiest job of the twenty-first century' [1b] and 'precision medicine' promising to provide 'the right patient with the right dose of the right drug, at the right time'. The 'value' derived from big data insights in general is ubiquitous in our lives with such data becoming ever more prevalent in businesses, with suggestions for new purchases and real-time alerts for possible credit card fraud, with applications from digital agriculture to personalized shopping experiences [2,3]. Similarly, precision medicine success stories in cancer and rare genetic diseases have transformed lives [4,5]. Combining these two approaches thus holds great promise for the future of medicine. Fulfilling this promise relies on the underlying assumption that these data can be transformed into knowledge by taking advantage of advances in technology (computing, artificial intelligence, genomics, sensors, etc.) to provide meaningful and actionable insights to guide medicine and drug discovery [6].

'Big data' is constantly being redefined with daily increases in the amount and quality of information generated. Ninety percent of the data ever created by the human race is estimated to have been generated over the past 2 years, and it is anticipated that the 'Internet of Things' will generate data volumes that double every 12 h rather than every 12 months, as is the case now. The pace of data generation is increasing across many different domains; by 2025, genomics, which will play an important role in healthcare, will generate some of the highest volumes [7]. These increases in data generation require equally important capacity for data acquisition, management, access, storage, and analytics; however, this cyber-infrastructure has not progressed at nearly as rapid a pace – indeed, Hu, et al. [8] note that big data research is as yet 'embryonic'.

Characterization of big data focuses on the 'Vs': volume, velocity, variety, veracity or validity, and value. These reflect not only the technical issues surrounding acquisition, access, transfer, and storage, but also those around heterogeneity and quality. The technical challenges are already being addressed in large companies such as Google, Facebook, and Amazon, and many of the insights garnered in the business realm will transfer readily to healthcare through the use of cloud computing resources, such as Amazon Web Services, Google Clouds, and Aliyun (a subsidiary of Alibaba). Challenges surrounding heterogeneity and quality present more difficulties given the variability in potential data sources, from the microscopic scale to the macroscopic, over time, across space, and in numerous formats (Table 1). In addition, critical challenges need to be addressed with regards to privacy and confidentiality issues that go beyond cyber-security; how the Institutional Review Board (IRB) and Informed Consent protocols are designed and implemented will determine how data is collected, stored, and accessed.

The unprecedented depth, breadth, and scale of data, which can now be collected drove President Obama's proposal of the

CONTACT Gunaretnam Rajagopal ✉ grajagop@its.jnj.com Vice President of Computational Sciences, Discovery Sciences, Janssen Research and Development LLC, 1400 McKean Road, Spring House, Pennsylvania, 19477, USA

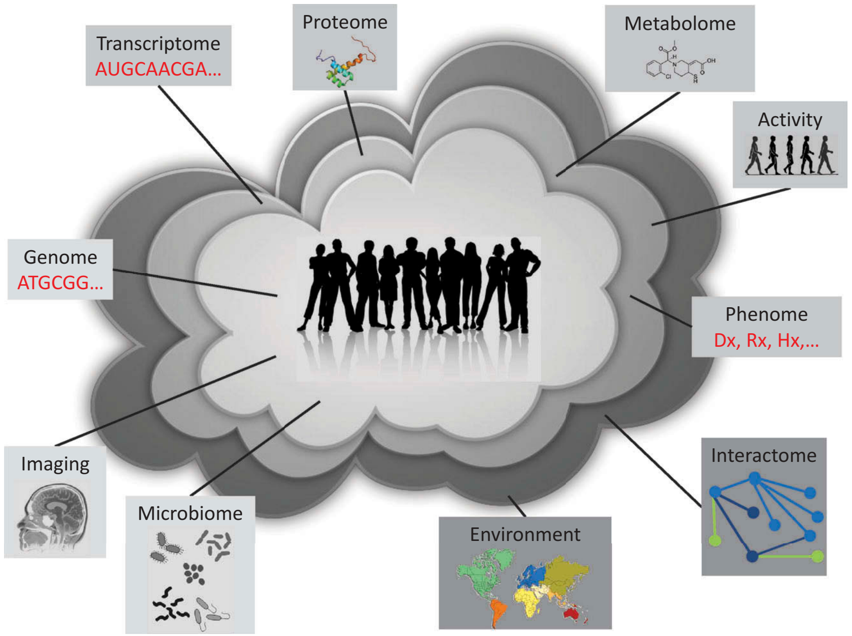

Figure 1. In the future, precision medicine will be enabled by individuals surrounded by a cloud of data. Layers of the cloud represent data varying between those directly affecting/quantifying the individual, and those quantifying the environment and indirectly affecting the individual, i.e. levels of the exposome.

Table 1. Characteristics of relevant big data types for precision medicine.

Data Source	Characteristics	Maturity
Genomics	Stable, structured volume	Innovative
'Omics	Time-dependent, structured volume/variety	Standardized/ optimized
Microbiome	Time-dependent, structured volume/veracity	Controlled
EHR, claims data, registries	Time-dependent, (un) structured volume/variety/veracity	Standardized
Imaging	Time-dependent, structured volume	Controlled/ standardized
Wearables/sensors	Time-dependent, (un) structured volume/velocity/variety/ veracity	Basic

Maturity levels [172].
Basic: disjointed, manual infrastructure, knowledge not shared, reactive and ad hoc.
Controlled: coordinated; manual infrastructure; knowledge silos; problem-driven.
Standardized – standardized infrastructure; individual-level collaboration; request-driven.
Optimized – consolidated and virtualized infrastructure; team-level collaboration; service-driven.
Innovative – strategic asset; enterprise-level sharing and collaboration; value-driven.

Precision Medicine Initiative in January 2015 [9] and will eventually enable the aims of precision medicine by surrounding individuals with a cloud of data (Figure 1). Big data generated through new technology developments in genome sequencing, 'omics platforms, electronic health records (EHR), imaging, sensors, etc., have the potential to transform the process of drug discovery and the delivery of healthcare.

Challenges facing big data in precision medicine

Incorporation of big data into everyday healthcare use can be guided by successes in other domains. Examples of successful analytics pipelines abound in everyday life, from Google Maps directing drivers in real-time toward the best route home to social network websites suggesting new things you might like or people you may know. In addition to overcoming the challenges of generation, acquisition, storage, and analytics, these examples also take the final steps of translating the results of algorithms into actionable insight, and making those insights useable by the general public. In healthcare, we believe formal methods of 'evidence synthesis' will play a critical role. This generally involves identifying and collating all relevant data; assessing bias; and assessing what weights (with regards to accuracy and reliability) to give different sources of data, depending on the question being asked [10]. Numerous foundational requirements are vital to the success of precision medicine implementation [11], which ultimately will be driven by the practical 'value' it provides to clinicians and patients, which in turn determines adoption and sustainability [12].

The successes of big data have spurred the uptake of big data analytics in healthcare, but it is easy to overlook differences in situation, which will challenge its implementation in precision medicine. A mistake from Google Maps may lead to a wrong turn and 10-min delay; even at a frequency of 20%, the benefit outweighs the risk. Mistakes in algorithms predicting disease risk, however, have costs measured in invasive procedures, money, time, and potentially life-changing and irreversible decisions, such as undergoing unnecessary or inappropriate treatment. Such outcomes are unacceptable at that frequency. Hence, much higher standards are required for big data in precision medicine. These standards will have to be supported by accurate and precise analyses based on high-dimensional biological data, which is frequently riddled with biases due to self-reporting, noise, and missing data [13]. We focus on major data sources (Table 1) relevant to precision medicine below, highlighting the issues specific to each source and requirements to bring them together coherently.

In common to these technologies is the need to answer three questions – what is the specific question/hypothesis being asked/tested; what data must be collected to answer it; and what quality of data will ensure that signal can be differentiated from noise, avoiding inconclusive or misleading 'answers'. As technologies improve, we can measure phenotype and environment in the normal and disease states on both small and large scales, spatially and temporally, from single cell sequencing [14,15] to global sensor networks [16]; from millisecond scale [17] to over decades [18]. However, much of the data collected is redundant and non-informative, and ideally does not need to be retained. We can learn from astronomy here, where 7.5 TB of data are streamed every second in the Australian Square Kilometer Array Pathfinder Project [7], but only very specific signals of interest are stored. Until the scope and granularity of measurement has been determined, we either run the risk of overloading systems and increasing cost (e.g. storing all sensor data), or discarding important information (e.g. ignoring rare variants in genomics). We do not expect initial studies to have this information at hand, but as we refine our understanding of disease, the question at hand will help guide the design, collection and analysis of big data.

We believe that big data in biomedicine has tremendous potential to transform healthcare, drug discovery and development. Nevertheless, we will highlight key barriers to bringing this potential to reality. For instance, relatively little attention has been given to find more efficient and sustainable financial and operational models for collecting, storing, organizing and accessing biomedical big data. This is especially true when the resources, cyber-infrastructure, data management, and access control process differ greatly in the research realm versus that within the clinical environment. Especially within the clinical space, meaningful insights gained by translating patient data to actionable, prescribing decisions necessitates robust, reproducible analytical tools easily accessible to providers – a challenge since the science, technology, and analytics are still evolving. We will address these issues in the following relevant sections.

'Omics data

The hope to transform the practice of medicine through data began with the Human Genome Project; though that was completed in 2003, the transformation is still in its infancy 12 years later [19–22]. This historical emphasis on genomics in medicine, with technologies such as transcriptomics, proteomics, and metabolomics (Figure 1) only being investigated much more recently, has meant that genomic computational infrastructure and analytical tools are the most mature and well developed. Falling costs, improved accuracy, and a rapidly expanding scientific knowledge base have brought genome sequencing to the cusp of clinical practice [23]. Nevertheless, there are still numerous issues in large-scale studies [24] due to rapid changes in technology from single nucleotide polymorphism (SNP) chips to whole exome sequencing (WES), to whole genome sequencing (WGS) as costs have decreased and concurrently the data generated have massively increased. Indeed, a major focus of the first phase of the Electronic Medical Records and Genomics (eMERGE) network was to integrate genomic data from multiple sites [25,26], addressing issues such as missing data and quality control.

Genomewide association studies (GWAS) represent the first major efforts to connect big genomic data to phenotype and to make medicine more precise through the detection of variants, which impact disease [27] and drug metabolism [28]. Standard approaches rely on statistical models, such as logistic regression and linear mixed models [29–31] with emphasis on correcting for potential biases and confounding factors and accommodating the scale of the data. Machine learning approaches for genetics and genomics (reviewed in [32]) also show promise for tasks, such as annotating genes, predicting function from sequence, and distinguishing between disease phenotypes. While GWAS have had great success in the sense of detecting thousands of variants associated with diseases and traits [33,34], many of these variants were found in non-coding regions of the genome and relatively few have been translated into therapies and/or diagnostics. In addition, for common chronic diseases, even hundreds of variants may only explain a small proportion of the overall variance attributable to genetics (heritability). In spite of this, Visscher et al. argue that GWAS have not been a failure, but rather that initial expectations were uninformed and unreasonable [27].

In retrospect, the idea that only information derived from the genome was sufficient for precision medicine seems naïve, though three areas (cancer, pharmacogenomics, and Mendelian disorders) with potential for short-term benefits rely primarily on genomics [35]. Recognition of the opportunities in cancer genomics and pharmacogenomics is evident by their inclusion as focus areas in the NIH-funded Precision Medicine Initiative [36], while the advent of non-invasive genetic testing methods has prompted great uptake in prenatal screening for Mendelian disease [37]. However, complex diseases are far less tractable. The question of 'missing heritability', the genetic variability which cannot be explained by identified variants [38], demonstrates the difficulty in predicting phenotype from a small set of actionable genotypes [35]. Models which incorporate information from the whole genome more fully explain the genetic variability [39,40], but these are more difficult to translate into action. It is anticipated that recent initiatives, such as the NIH-funded ClinVar Consortia [41,42] will help address this.

Another flawed initial expectation was the underlying 'common disease common variant' hypothesis. Until sequencing became sufficiently cost-effective, only common variants were tested in SNP-chip GWAS. Since then, however, it has become clear that (rare) disease variants under negative selection pressure and hence observed at very low frequencies in populations (minor allele frequency less than 1%) can have large effects on complex traits [43–45]. However, rare variants do not fully explain genomic variability and the complexity induced in phenotype.

The growing recognition of the complexity of the genotype–phenotype connection has led to consideration of the full range of biological variability (systems biology), from genomics and its interactions to modifications of DNA (epigenetics), post-transcription (transcriptomics), post-translation (proteomics), post-metabolism (metabolomics), to non-human (microbiome/metagenomics) (Table 1). The system varies with time and location of the sample taken from the body, so big data generation and analytics present more challenges than in the context of the genome. However, the increased and dynamic information has the potential to thoroughly characterize changes associated with disease [46,47]. Studies are now being initiated to scale these approaches to the population-level [48,49], though it is not yet clear whether the insights derived will justify such extensive testing. These studies must contend with issues arising within data type regarding data quality, filtering, dimension reduction, and heterogeneity, as well as the corresponding issues related to integrating the diverse data sources [50]. A particular concern will be the integration of data across different levels of maturity – newer approaches, such as the microbiome still suffer from a lack of consistency across studies, in protocols and analysis, which is only starting to be addressed [51], whereas standards for RNA microarray studies have been required at most journals, since 2001 [52].

EHR/claims data

Electronic health records have been in use in the United States, since the 1960s and 1970s, although even as late as 2009 were only implemented for 10–20% of doctors and hospitals [53]. The Health Information Technology for Economic and Clinical Health Act then instigated rapid changes by offering a combination of incentives and penalties for EHR adoption (and lack thereof) by 2015. As of 2015, the rate of adoption is nearly 80% in the United States, with varying rates in other countries [54]. The Scandinavian countries have generally led the way; the Danish National Patient Registry has covered all hospital encounters for the entire population since 1996, containing over 65 million total clinical encounters [55].

While EHR databases provide rich resources to drive precision medicine development, it cannot be forgotten that this is but a secondary use of the data. The data accumulated into EHR flows naturally from clinical processes; however, it is constrained by the EHR format, which presents novel challenges for standardization and curation. Structured and unstructured data are stored on billing and diagnoses codes, electronic prescriptions, laboratory reports, narrative text notes, and are gradually being extended to incorporate imaging and genetic data as well [56–58]. Entry of this data varies across many levels: across vendors, who may have individual storage formats; within vendor across systems (for example, supporting the different needs of pediatric oncology versus radiation oncology), which may have specific customizations; and within system across individuals, who may enter unstructured text in different ways. This fragmentation is further exacerbated in the United States and other systems without a national ‘from cradle to grave’ health service, where individuals may have EHR in several systems during their lifetimes. Data processing must thus accommodate all of these variations, in addition to making use of natural language processing and text mining [58–60] to pull out hidden information from records.

Even assuming standardized, structured data, analysis is not a simple task. All data are observational, and hence subject to classical epidemiological problems, such as confounding and bias in sampling, in the data capture process, and in the health-care delivery system. Further, while observations are taken over time, they are taken opportunistically, making it difficult to disentangle cause and effect; side effects from medications vs. new indications; and various dependencies [61]. Issues of spurious correlations and incidental endogeneity are highly relevant in EHR [13]; highlights methods for dealing with these. Such problems will only grow as large and distributed EHR data are merged for population-based studies.

EHR implementation has been motivated by the promise of clinical benefits, including minimizing human errors, maximizing cost-efficiency, and increasing coordination between centers. Hence, systems have been designed with these goals in mind, although Kukafka et al. [62] provide suggestions for how to better align EHR with public health goals. Rushes to implementation can and have adversely impacted final value derived from these systems; implementation should be considered carefully in future integration of EHR with data sources, such as genomics and imaging.

In recognition of this need for robust and rigorous analyses with such large datasets, many efforts have been made by the research community to provide open-source software tools for EHR analysis. One focus area has been the development of robust and validated phenotyping algorithms [59,62–64], many of which are publically available at [65]. Methods for other tasks, such as patient-level prediction, population-level estimation, study population evaluation, and safety surveillance are being developed by public–private partnerships, such as OHDSI [66] as well as the FDA [67]. These tools are not currently adapted to deal with sources of data, such as genomics and imaging, but provide an excellent starting point for future integration.

Imaging data

Images have formed an integral part of the health care system for over a century, with X-rays first being used in medicine only a year after their discovery in 1895 by Wilhelm Röntgen [68]. In many ways, numeric image data, represented by intensity measurements for each pixel, are similar to genomics data. They form a set of high-dimensional, highly correlated variables, with high-resolution magnetic resonance imaging (MRI) and computed tomography (CT) scans easily in the gigabyte size range. Given the size of these data, standards are important – the digital imaging and communications in medicine (DICOM) standard is a critical component of any radiology department in order to integrate modes and vendor formats [69].

A major difference between genomics and imaging is their interpretability by clinicians. For diagnosis, the visual nature of images is straightforward compared to the unintelligible string of three billion letters formed by a whole genome. In that sense, a picture is worth much more than a thousand 'words'. This results in a bottleneck at the clinic, however – automated analysis and inclusion in the electronic health record is even now rare, in spite of the importance of images for many specialties [70]. Biomedical imaging informatics has arisen as a subspecialty of biomedical informatics due to the critical need for developments in this area, and a special issue [71] reviews topics in this field covering image characterization and management, modeling and decision support, and applications to improve clinical practice.

Large imaging datasets are now being aggregated in many general imaging databases [72–74], including those with specific focus areas such as cardiology [75] and incorporating genomic data [76]. Indeed, the Cancer Imaging Archive [77], which is linked to The Cancer Genome Atlas [78], includes 31 sites and had over 3 million images uploaded in the first year of operation. The National Cancer Institute (NCI) further sponsors initiatives in quantitative imaging [79] which produce rich resources for many different diseases and imaging types. The value of these databases is also being recognized by industry; in 2015 IBM acquired access to 30 billion medical images to be utilized by Watson Health for deep learning [80].

Automation of analysis for this large and rich data source will allow development of predictive analytic tools based on high-throughput feature extraction [81]. This approach, 'radiomics', can extract descriptions of intensity distribution, spatial relationships, and heterogeneity in shape and texture to characterize areas of interest such as tumors. It is built on the foundation of computer-aided diagnosis using medical imaging, for which algorithms have been developed for organs including the brain [82]; prostate, breast, and lung [83]; and bone [84]. The brain and cancers represent major areas of interest for development of biomarkers due to the need for noninvasive diagnosis and monitoring techniques [85]. However, both biomarkers and diagnostic algorithms often suffer from issues with sensitivity, specificity, robustness, reliability, accuracy, and fragmentation of the data resources required. As with 'omics, the transition to regular use will require strong infrastructure, integration with other data types, and ultimately, validation of clinical utility.

Sensor data

The most recent addition to the medical data catalog comes from sensors, which have become ubiquitous with advances in computing and the spread of smartphones. Sensor data can essentially be streamed continuously, or at least at far higher throughput temporally than other data sources; in addition, it can cover a wide range of spatial locations, either through many fixed sensors, or mobile sensors with varying position over time. In theory they provide real-time, cheap, objective measurements of quantities which are otherwise unknown or only measured through more expensive in-depth clinical tests. Use cases include monitoring temperature, inhaler use, glucose, seizures, hearth rhythm, activity, sleep, stress, and sweat chemistry [86]. These technologies hence offer great promise, particularly for diagnosing and monitoring cardiovascular and metabolic diseases [87] and mental health [88,89].

In practice, sensors present numerous challenges, in part due to their novelty, in part due to the volume and velocity of the data they generate, and finally due to the need to 'normalize' data collected across different sensors to enable real-time accurate decisions and catalyze behavior change. While there are many options available through smartphones, watches, and specialized devices (e.g. FitBit, JawBone, etc.), they have relatively few features, and existing apps may have bugs and systematic biases. Little standardization or validation of measurements has been undertaken. Hence, in measurements of exercise levels, three different devices can store entries of active time, steps walked, and distance traveled. Exercise outside of walking, running, or cycling must be input manually, and frequently even the automated capture contains errors (for example, mistaking cycling for walking). These issues introduce an element of subjectivity and necessitate additional effort, which can lead to reduced adherence. Some, known as self-quantifiers, retain interest for long periods of time [90], but in a study following undergraduate students for 6 weeks, 65% dropped out in just over 2 weeks [91]; the value is simply not yet evident. The ability to collect data is not sufficient in itself; to drive uptake requires design and analysis features to facilitate behavioral changes [92]. In terms of having longevity, then, sensor development may require a move to devices which are unobtrusive, energy-efficient, and require little to no effort for data capture and transfer via wireless – implantables, potentially, rather than wearables. In terms of producing long-term valuable medical data, they will require the extensive validation, qualification, and regulation mandated for other medical devices to demonstrate the clear value added and correspondence with more expensive and burdensome tests.

The volume and velocity of data generated presents problems when considering data processing, storage, and interpretation. It is still unknown which features drive biological responses – whether gradual trends, sharp spikes, accumulation of events, or some combination thereof which will have clinical impact. Once evaluated, such knowledge would allow immense data compression, easing storage, transfer, analysis and interpretation. Until then, all individual data must be stored in order to build predictive models of disease risk, even if most are uninformative or redundant. Data mining and machine learning techniques have much to offer in terms of tasks such as anomaly detection, prediction, and classification [93,94] but must accommodate systematic biases in device use. Ultimately, we expect to see a range of complementary devices available, from low-cost sensors such as those in smartphones, to high-precision instruments for specific focus areas. This will allow decreased accuracy and precision of individual sensors to be balanced against overall information generation.

Data analysis in mobile health (mHealth) will be facilitated by careful design of devices and by collaborations across domains with longstanding experience in spatial and temporal data. A key issue will be identification, testing, and qualification of biomarkers, which will require not only robustness and reliability, but also evidence of meaningful relationships with other, clinical measurements [85]. As for EHR, major risks include the detection of spurious correlations and incidental endogeneity [13]. This nascent area will benefit from developments in fields such as econometrics, where identification of trends and predicting future events forms the backbone of the finance industry [95]. Other domains, such as molecular simulation, environmental and biological analysis, also have methods to offer, as in the area of detection of change points across several time series [96]. Indeed, spatial sensor data, while only just beginning to be used in healthcare, has a long history in environmental and agricultural studies, where climate and rainfall are of direct interest [18]. The integration of sensors and modeling for environmental variables, such as air quality, climate information, population movement networks, and personal time-activity patterns will ultimately have great impact on human healthcare [97].

Data integration

Standardizing, integrating, and unifying data is challenging both within and across data types, but is absolutely crucial to capitalize on the opportunities presented by big data in precision medicine. Examples of genetics adding little value beyond clinical utility [35] in traits, such as type 2 diabetes, stroke, and rheumatoid arthritis in part reflect incompleteness and noise in the data collected from underlying phenotypic heterogeneity; more precise definitions of disease from registries and EHR may hence improve the power of genomic approaches [98]. The linkage of EHR with genomic data was initiated primarily through the eMERGE network [99], but is now having widespread uptake throughout academic and industry partnerships [100,101]. Starting from genomics, EHR can be used for validation [102], or conversely can provide insight into novel subtypes of disease through dissection of patient similarities, and then GWAS can identify genetic variants underlying these subtypes [103]. One challenge here is the differential in quality between genotype data, which has improved tremendously, and the phenotypic data, which are highly variable in quality of information recorded (touching on the issues of veracity and reliability in big data).

Including sensor and imaging data will expand both the degree of phenotypic characterization and the number of potential explanatory variables. A major focus with images is to identify novel endophenotypes, intermediate traits, which may be more closely related to genomics than disease status; a major focus with sensors is instead to better characterize the environmental and lifestyle factors which complement genomic data. Thus, GWAS methods for testing association between responses and factors accommodate both types of data, although their variability over time and space, and high-dimensionality may require multiple stages of modelling.

GWAS conducted with imaging traits highlight the need for feature identification and sparse testing methods to reduce the dimension of the phenotype space. The Alzheimer's disease neuroimaging initiative (ADNI) has identified variants associated with specific features in targeted regions of the brain and structural phenotypes derived from the images, as well as deriving statistical maps of the whole brain from voxel-based GWAS [104]. For the most part these approaches rely on brute force testing, which has little power given the vast number of potential associations (tens of thousands of voxels with hundreds of thousands to millions of SNPs [105]). By first identifying structural regions associated with the disease of interest (e.g. Alzheimer's) rather than testing for associations directly with disease, trait heterogeneity is reduced, increasing power and narrowing the biological focus; however, the extra stage then linking results back to disease status may introduce additional uncertainty in analysis. Expanding the model further, Hyde et al. [106] make the case for integrating imaging genetics with the environment. Such analyses frequently suffer from lack of power given the complexity of gene-environment interactions, the large number of tests involved, and the reduced sample sizes compared to studies of height or more common and easily measured traits. In such studies, therefore, endophenotypes become even more important.

Analysis of environmental interactions will also benefit from the in-depth characterization made possible by sensors. Previous gene-environment analyses have focused primarily on classical epidemiological exposures such as tobacco, diet, and chemicals. However, the full 'exposome' [107] includes the general external environment (e.g. climate, education, urban environment), the specific external, which are typical epidemiological exposures, and the internal (e.g. microbiome, inflammation, metabolism). Much of the external environment can now be profiled in a high-throughput, time- and space-varying objective manner using sensors, including the use of 'omics technologies to characterize the internal environment and its response to external exposures [108]. However, the combination of these data sources requires new developments in big data infrastructure and analysis [109].

Integrated models are not yet sufficiently developed to where they can accommodate both the diversity of data sources, but also their dimension and speed. Agricultural and environmental applications have used Bayesian modelling for sensor data and high-dimensional datasets varying in time (e.g. imaging, 'omics, etc.) [110,111]; however, the implementation of these models is not yet fast enough for real-time updates, particularly once genomic and EHR data are also included. With more mature machine learning techniques such as dimension reduction analysis of sensor data, however, there is promise for models which update on a weekly-to-monthly time scale, which is likely fast enough for most diseases. In the recent 'MyConnectome' study [112], fMRI time series on a healthy individual were integrated with transcriptome and metabolome data collected over the course of 532 days at weekly-to-monthly intervals. This publically available data will provide rich information to improve designs and analysis approaches for future studies incorporating imaging, genomic, and behavioral data. For individuals already diagnosed, a tiered approach may become popular, with specific models reflecting the size of the population at risk and the speed of disease progression, as well as the relevant risk factors.

Oncology: success story or still a major challenge?

For more concrete examples of using big data to enable precision medicine, we turn to oncology, which has a strong history in the area (detailed in [113–115]). Well-defined driver mutations (e.g. BRAF [116,117]; KRAS [118]; EGFR [119–121]) have long formed the basis of targeted treatments. While few of these were discovered using big data approaches, sequence analysis of both tumors and individuals is now common to match patient to treatment. Genomics and gene expression have already been used to develop novel classification systems for lung, breast and prostate cancer [122–124]. Imaging and EHR data are regularly collected to track disease progression, and are now being aggregated and analyzed to improve the quality of care (e.g. [125]). Indeed, clinical trials have already been designed around precision medicine strategies, integrating both bioinformatics and clinical data to deliver real-time input to decision-making (e.g. [126]).

Many approaches for big data in precision medicine can be framed as exploiting the similarities in a homogeneous subset to predict the course of disease. In cancer, this is required by the spatial and temporal heterogeneity of tumors [127,128], which both affects and is affected by treatment. The consequences of their fine-grain evolution have motivated prospective trials to track and resequence large cohorts after relapse [129]; investigation of the tumor microenvironment [130,131]; and collection of data from single-cell sequencing [14,15]; all of which have analogies for other diseases. These strategies face the challenges we have described previously related to the volume and veracity of data from new technologies and integrating longitudinal ‘omics, imaging, clinical, and environmental information.

Not all of the successes in cancer will translate generally; however, the lessons learned in developing resources and strategies provide a roadmap. A major factor in success will be disease etiology. In particular, cancers are often affected by somatic mutations with large effect sizes, while many complex traits are primarily affected by germline mutations with smaller effects. This may require different strategies, larger samples, and more complex models to develop effective targeted treatments. In turn, this will require widespread community efforts to collect big data for thorough characterization of the disease in large populations [77,78,125], enabling the definition of more homogeneous disease subtypes [124], and eventually leading to adoption of novel forms of clinical trials [132].

Ethical, legal, and social implications

Accompanying the many technical issues are the ethical, legal, and social implications of a shift toward the *practice* of precision medicine. These include general topics, such as socioeconomic and health disparities; required behavioral, economic, and legal changes; and interoperability, data sharing, privacy, and confidentiality concerns. Data sharing and handling regulations vary widely within and across geographical boundaries. To address this, a multinational coalition (the Global Alliance for Genomics and Health [133]) has developed a framework consisting of guidelines on privacy and consent, accountability, etc. In the specific context of EHRs, many of these issues are subject to the Health Insurance Portability and Accountability Act (HIPAA) regulatory framework [134] and are reviewed by [135]; however, not all of the unintended consequences of EHR implementation have yet been resolved [136]. In particular, the issues of data sharing, information overload, lack of clinical decision support, and genomic literacy among patients and care providers will only be exacerbated with the integration of additional data sources. While appropriate automated and standardized systems are clearly necessary to reduce the data flow to manageable levels, care is essential in implementing, maintaining, and updating novel precision medicine systems; unintended and unforeseen consequences can undermine safety and quality, and ultimately lead to a lack of uptake.

The adoption of EHR in countries around the world provides valuable lessons [54], but with new types of data will come new challenges. Mobile-health applications are likely to be more open to security breaches than EHR; genomics and imaging are more high-throughput and intensify hurdles related to data storage, transfer and interpretation; the rapidly evolving nature of our understanding of biology raises the issue not only of what genomic information is actionable, but also how often it must be reviewed in the context of new information; ultimately, these bring additional concerns related to liability for health care providers in prescribing appropriate care to the fore. It is hoped that inferences drawn from appropriate analysis of big data collected within the healthcare ecosystem will go a long way to mitigate this, though caution is advised to avoid ethical violations in the use of big data analytics. To address this concern, several federal agencies, including the Department of Health and Human Services, are proposing amendments to the common rule governing the ethical conduct of research involving humans. A significant change is the requirement of one-time, general, and open-ended consent for the use of all bio-specimens (and associated derived data), whether or not they are de-identified, for broad research purposes in the subsequent 10 years [137]. While this will enable much more extensive sharing and usage of data, care in implementation is required to alleviate privacy concerns which could adversely affect patient engagement.

Addressing these challenges relies on stakeholder engagement and filling education gaps across a wide spectrum of organizations. The eMERGE network has proved a testing ground in engaging groups ranging from organizational leadership, to clinicians, patients, research institutions, government, developers, and health-payers [138]. Popular support from patients and providers is critical to achieving precision medicine initiatives, for sharing data as well as participating in the research which will drive advances in learning health systems, predictive models for disease, and development of novel therapeutics. Fears about security, privacy, and confidentiality must be addressed, not only through appropriate regulation, laws, and standards, but also through education on the importance of data, the need for it to be shared, and the likelihood and consequences of security breaches [139]. In fact, one use of big data would be in training and educating providers and patients with real world evidence at the levels of the individual and the population to better understand disease and response to treatment. The need for education does not stop at the consumer, however. Clinicians currently are neither trained to deal with the volume of data now available, nor with systems which may help them manage it. The demand for multidisciplinary occupations, such as genetic counselor and biomedical informaticist demonstrate the need to fill gaps in the system with roles bridging data and interpretation. Training initiatives now being implemented in traditional university degrees, society-sponsored programs [140], and massive open online courses [141,142] will help to address this demand in coming years.

New roles in a disrupted healthcare system

Ultimately, the entire healthcare system will feel the impact of disruption from big data, from patients and clinicians to academia, industry, regulators, and health-payers. The need for participation from all relevant stakeholders, across domains and international boundaries, has begun to be addressed by large national and international consortia, but this level of cooperation must grow to realize a successful implementation of universally accessible precision medicine. In addition, we expect to see the rise of new business models for digital medicine which combine facets of biopharmaceutical, medical device, and technology to remotely monitor patients and collect data; aggregate, analyze, and sell that data; and translate insights from those data into actionable real-life interventions for high-risk individuals [143].

The stakeholders for precision medicine can be divided into four broad categories with distinct roles. The first contains government and large companies, such as big pharma, and technology drivers, such as Google, Microsoft, and Apple among others (see Box 1). Natural roles for this group are regulation, development of infrastructure, scaling and streamlining processes, and generally ensuring standardization and harmonization of efforts by coordinating centralized efforts and bringing parties together [144]. This group operates at a high level to catalyze change, with a global presence, working across and in cooperation with multiple bodies from other groups. Stakeholders are hence well suited to overcoming issues with translation and implementation in order to sustain the big data ecosystem necessary for precision medicine [145]. Activities might include establishing metrics for data usage in order to determine what to retain and maintain; offering incentives for curation and annotation; developing business models for data access; and promoting the information commons (e.g. Box 2). Both public and private entities have already initiated efforts to streamline and harmonize the field by providing tools to the general research community [146,147].

The second group is composed of smaller stakeholders such as academic groups and technology, biotech, and device startups. They are less constrained by complex internal processes governing data access and control (given their relatively simpler organizational structure) as well as bureaucracy, and hence more agile and willing to take on risk by testing novel ideas before passing them on to the first group for further development. However, their smaller size constrains the resources that can be employed and makes coordination more difficult unless larger consortia are formed. Private–public partnerships such as

Box 1: Large population-based studies/cohorts with recruitment occurring over next 5 years

We describe briefly several major private and public initiatives [148] to recruit, genotype, and collect big data on large population cohorts with the eventual aim of guiding precision medicine.

23andMe [149]
Direct-to-consumer DNA collection, genotyping and analysis service with large collection of survey data attached to genotypes of over one million individuals, with most consented to use data for research purposes.

Apple ResearchKit [150]
Open source software tool for medical researchers, doctors and scientists to collect and monitor patient data through apps for the iPhone and Apple Watch.

BGI 3 Million Genomes Project [151]
Plan to sequence one million individuals (as well as one million plants and animals and one million micro-ecosystems). Over 50,000 sequenced with theoretical capability to sequence one million/yr.

deCODE Genetics [152]
Leverages population and family-based resources to discover genetic risk factors for common diseases, create new means of diagnosing, treating and preventing diseases. Research is focused on 140,000 participants in Iceland (>50% population); currently a subsidiary of Amgen.

Genomics England – 100 K Genomes Project [153]
UK Department of Health-backed sequencing of 100,000 genomes by 2017, with focus on cancer (25 K patients; 25 K tumors) and rare disease (15 K patients; 35 K healthy relatives).

Human Longevity Institute [154]
Genomics and cell therapy-based diagnostic and therapeutic company to tackle aging-related diseases, with goal of sequencing one million genomes by 2020 through private efforts. Capability to sequence 40 K genomes/year.

Institute of Systems Biology 100 K Wellness Project [155,156]
Expansion of pilot Hundred Person Wellness Project capturing information on sleep patterns, proteomics, immune cell activity, pulse, physical activity, microbiome, blood glucose, and WGS. The project has been spun out into Arivale, a new scientific wellness company.

NIH Precision Medicine Initiative [9]
National cohort study of one million or more Americans recruited by 2019, collecting specimens, genomics, EHR and lifestyle data. Initial focus on oncology, rare diseases, and pharmacogenomics.

Verily Baseline Study [157]
Initial pilot study of 175 volunteers expanded to larger multicenter collaboration with Duke and Stanford to collect and analyze genomic, wearable, molecular, and imaging big data from 10,000 volunteers to understand risk factors and progression of cancer and cardiovascular disease.

Box 2: NIH Big Data to Knowledge (BD2K) Initiative [158]

This initiative funds research and training activities, including methods and tools development, to support the use of big data in biomedical research and discovery. Major efforts are funded through the BD2 K Centers of Excellence, of which 12 have so far been established (briefly described below).

LINCS-BD2 K Perturbation Data Coordination and Integration (Icahn School of Medicine at Mt. Sinai)
Aims to produce novel analysis tools and methods to translate insights from biological model systems into drugs and pathways for complex diseases.

Big Data for Discovery Science (University of Southern California)
Development of methodologies, software, and interactive visualization tools for efficient large-scale analysis of proteomics, genomics and images of cells and brains.

Big Data in Translational Genomics (University of California Santa Cruz)
Development of data models and analysis tools through multinational collaboration between academia and industry, with initial focus on gene variants contributing to cancer.

Causal Modeling and Discovery of Biomedical Knowledge from Big Data (University of Pittsburgh)
Emphasis on Bayesian tools for causal models with focus on cell signaling for cancer; susceptibility to lung disease; and functional connections in the human brain.

Expanded Data Annotation and Retrieval (Stanford)
Facilitate automated annotation of data by developing metadata standards and repositories.

Predictive Computational Phenotyping (University of Wisconsin-Madison)
Predict disease risk using machine learning and statistics, by integrating data including EHR, imaging, and molecular profiles for breast cancer, heart attacks, and severe blood clots.

Mobile Sensor Data-to-Knowledge (University of Memphis)
Facilitate gathering, analyzing and interpreting data from mobile sensors in the context of reducing hospital readmissions for congestive heart failure and preventing relapse from smoking cessation.

A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform (UCLA)
Democratize data research with community-driven data integration and modeling for cardiovascular research in the study of protein structure, function and networks.

ENIGMA Center for Worldwide Medicine, Imaging and Genomics (University of Southern California)
Develop computational methods for integration, clustering, and learning from complex and diverse biomedical datasets in a global effort to combat human brain diseases.

KnowEng, a Scalable Knowledge Engine for Large-Scale Genomic Data (University of Illinois Urbana-Champaign)
Leverage data mining and machine learning techniques through a knowledge engine to combine gene function and interaction information from diverse genomic data sources.

Mobility Data Integration to Insight (Stanford)
Provide access to mobility data for over ten million people, along with tools to improve outcomes in the areas of gait pathologies, movement impairments, and increases exercise for joint health.

Patient-Centered Information Commons: Standardized Unification of Research Elements (Harvard)
Develop systems to integrate genetic, environmental, imaging, behavioral, and clinical data from multiple sources.

the IMI [159] or eMERGE network can offer a bridge between the first and second groups.

The third group is composed of health care providers and payers. In the United States, the Office of the National Coordinator for Health Information Technology is a key stakeholder [160], whose mandate is ‘to support the adoption of health information technology and the promotion of nationwide health information exchange to improve health care’. In addition, American healthcare providers have been working together to address many policy issues relevant to reducing health disparities and to improve the quality of care through the use of big data [161]. While this group is less actively involved in the development of methods for precision medicine than the other two groups, they are perhaps the most heavily invested in its success.

The final group consists of not-for-profit foundations and patient advocacy groups. Examples include the Simons Foundation Center for Data Analysis [162], whose mission is ‘to create new computational frameworks that will enable scientists to analyze the large and complex datasets that are being generated by new experimental technologies’ and the American Society for Clinical Oncology [163], who have actively deployed tools such as CancerLinQ based on analysis of EHR data from thousands of oncology practices, to assist oncologists in effectively integrating hereditary cancer risk assessment into practice. They and sister organizations, such as Stand Up To Cancer [164] will play the role of ‘honest brokers’ in galvanizing all the relevant stakeholders to address many of the sociopolitical and technical challenges highlighted previously.

What could derail the progress of precision medicine?

Visions for a big data-enabled precision medicine future often center on a secure, seamless integration of patient data

Table 2. Characteristics which must change in order to achieve precision medicine goals, relevant data sources, and opinions on how to get there.

		Data source						
Present	Future	Genomics	-Omics	Microbiome	EHR	Imaging	Sensors	Getting there
Specific consent	General consent, ubiquitous biobank with metadata and adaptive clinical trials	✓	✓	✓	✓	✓	✓	Greater patient engagement; flexible data management to allow different sharing levels, enhanced cybersecurity
Population trials	n-of-1 trials	✓	✓	✓	✓	✓	✓	Critical mass of n-of-1 trials and infrastructure [168]
Big data infrastructure for business	Big data infrastructure for healthcare	✓	✓	✓	✓	✓	✓	Vendors with proven success in healthcare; collaborations between technology and healthcare industries
Widespread data generation adoption	Widespread data analytics adoption	✓	✓	✓	✓	✓	✓	Understanding of how to add value with analytics; understanding of what needs to be measured; better understanding of disease
Binary disease concept	Probabilistic disease risk framework	✓	✓	✓	✓	✓	✓	Education and training of healthcare providers and patients
Clinicians diagnose and treat	With aid of learning healthcare systems (LHS)	✓	✓	✓	✓	✓	✓	Development of LHS; role redefinition; education
Competition among vendors, researchers, institutes produces fragmented data and protocols	Harmonization of data standards and procedures	✓	✓	✓	✓		✓	Community-wide efforts to quantify variation and its effects in order to agree on standards
Rapid changes in technologies	Stabilized low cost, high-throughput technologies		✓	✓			✓	Attainment of threshold in resolution/cost with low error rate sufficient to justify large-scale implementation
Quantification of self	Effortlessly, and can be interpreted meaningfully					✓	✓	Devices become unobtrusive; Behavioral shift; evidence of value for health
Thousands of untested apps	Smaller number of tested, regulated apps						✓	Regulation guidelines; demonstration of value in healthcare to consumer
Basis for regulation and payment is *outcomes*	Basis for regulation and payment is *value*				✓			Definition of value; real-world evidence collection and analytics to quantify value

collection and clinical care, ubiquitous entry into clinical trials and Biobanks, and continuously learning health care systems. Table 2 proposes some characteristics of the status quo, which must evolve to reach this goal, but progress will take time. One of the biggest hazards, common for new technologies and paradigms, will be confusion over jurisdiction and other regulatory issues. Already, disputes such as that involving the FDA and 23andMe [165,166] have incurred much discussion over the role of the private sector in providing direct-to-consumer testing. Such controversies are typical signs of disruptive innovation, which challenges established interests and paradigms to motivate accommodation of new advances [167]. Regulatory bodies must balance their need to protect the interests of consumers against constraints which may hamper such innovation.

Indeed, realization of precision medicine will require widespread buy-in, critical both from the perspective of sharing cost (hence a need for a sustainable financial model) and risk, as well as ensuring harmonization and sustainability of resources developed. One can imagine a future with insufficient funding and capability to validate a vast array of apps, diagnostics, and devices. Erosion of confidence in the system leads patients to avoid sharing data and self-diagnose online; doctors are left to prescribe ‘at your own risk’; direct-to-consumer companies proliferate and provide incomplete or erroneous information. Indeed, the FDA has recently moved to regulate lab-developed tests more closely [169] amid concerns of a lack of evidence supporting clinical validity and deficient adverse event reporting. Their 20 case examples of diagnostics potentially causing harm to consumers are a cautionary tale against complete lack of regulation in parts of the medical space.

Moving too far in the opposite direction, of course, with a top–down approach and severe regulation, will only delay the advent of precision medicine. Getting the balance right will require good faith from all parties involved to avoid issues of indemnification with insufficient evidence, and to allow sufficient time for education and training necessary for the new era (e.g. genetic counselors, data scientists, etc.) to catch up with demand. Precision medicine is still in its infancy, but the vision is clear – what remains now is to ensure that the data infrastructure, quality control, analytics, training, and funding models are clarified and implemented in a coherent and collaborative manner to realize that vision.

Expert commentary

Big data is poised to have enormous impact on the healthcare ecosystem, with some estimates of its potential annual value being 300 billion dollars in the United States alone [170]. Data generation has been greatly facilitated by new technologies, but later steps in the data processing, analysis, and interpretation value chain have not seen rapid progress. Outstanding issues are primarily related to data integration and standardization, including combining ‘omics across time and tissues, normalizing and scaling data from different individuals, samples, and modalities, and identifying clinically valid data-driven features. Specific questions of interest will drive the design of studies and analysis of data for various diseases, though all must find ways to manage the complexity of biological processes. In addition to the technical challenges of big data, many sociopolitical challenges remain surrounding privacy, confidentiality, liability, financial sustainability and interpretation. While none of these challenges are simple to surmount, large cross-sector collaborations and widespread engagement will facilitate progress. Solidifying the entire big data infrastructure, filling gaps in the talent pipeline, and ultimately enabling physicians to draw on experiences and learning based on millions of individuals will bring about the era not just of precision medicine, but of P4 medicine – predictive, personalized, preventive, and participatory [171].

Five-year view

Over the next 5 years, both the quantity and quality of data collected for precision medicine will advance immeasurably. Million-person cohorts (see Box 1) will be recruited in multiple countries through both public and private endeavors, with individual genomic, environmental, and lifestyle factors recorded longitudinally. These cohorts will be valuable for investigating genetic diversity within populations, and gene–environment–lifestyle interactions, but to accurately predict risk for specific diseases, more focused studies will be necessary, potentially recruiting individuals from the larger cohorts. In these studies, by leveraging technological innovations and decreasing costs, longitudinal 'omics samples will be collected from relevant tissues more frequently than possible in a population study; reduction of scope will enhance resolution of measurement. Individual data analysis strategies and protocols will be standardized and scaled, with cloud computing becoming the norm rather than the exception; however, data integration will be a key issue. These advances will be enabled by increased training of data scientists, genetic counselors, and biomedical informaticists to narrow the talent gap.

Despite the rapid progress in resolving technical challenges, the sociopolitical framework underpinning healthcare will still be evolving. The introduction of learning health systems will follow a similar trajectory to EHR systems and will be widespread, but of varying quality across and within countries. Health payers will have moved to a value-based reimbursement model guided by big data and empowered by patient participation. While dominant business models for digital medicine will have emerged, regulation will still be in flux as companies remodel their operating structure to fit the new paradigms of precision health and P4 medicine. These developments will represent massive leaps forward in our understanding, prevention, and treatment of disease, but even so, we will not have unraveled the full complexity of biological processes, merely given ourselves the tools to eventually do so.

Acknowledgements

The authors thank members throughout Johnson & Johnson and external partners for discussions, scientific and technical collaboration in addressing the challenges of precision medicine and Big Data.

Financial and competing interests disclosure

The authors are employees of Janssen Research & Development and Johnson & Johnson Innovation (Johnson & Johnson compaines). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Key issues

- Big data and its analysis are rapidly becoming a mainstay in industry and have enormous potential benefits for healthcare by enabling the delivery of precision medicine.
- While stages of the big data pipeline (generation, acquisition, processing, storage, analysis, and interpretation) are well developed for business purposes, they present novel challenges in the context of healthcare.
- Genomics and other 'omics technologies, electronic health records, imaging, and sensors present novel sources of big data, varying over time and space.
- The relevance of data sources will depend on a deeper understanding of disease etiology to determine the level of resolution and frequency required in measurement; the question at hand should guide the design, collection, analysis and interpretation of data.
- The major outstanding technical challenges include data access and management, data standardization, data integration and validation.
- Oncology has achieved the most success in precision medicine to date; valuable lessons include widespread community collaboration and stratification to address heterogeneity in disease.
- Ethical, legal, and social challenges present as many issues as technical challenges, including data privacy, confidentiality, and security, as well as interpretation and implementation for clinicians and patients.
- The disruptive innovation in the practice of medicine induced by big data will result in new roles throughout the healthcare system, and require strong collaborations across public and private entities.

References

Papers of special note have been highlighted as:
• of interest
•• of considerable interest

1a. FDA. Paving the Way for Personalized Medicine: FDA's Role in the New Era of Medical Product Development; Available from: http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalizedMedicine/UCM372421.pdf; 1b. Data scientist: the sexiest job of the 21st century. Davenport TH, Patil DJ. Harvard Business Review, 2012. Available at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ [Last accessed 28 Dec 2015].

2. Morgan L. Big Data: 6 real life business cases; 2015 [Last accessed 17 Dec 2015]. Available from: http://www.informationweek.com/software/enterprise-applications/big-data-6-real-life-business-cases/d/d-id/1320590?image_number=1

3. Laskowski N. Ten analytics success stories in a nutshell; 2015 [Last accessed 17 Dec 2015]. Available from: http://searchcio.techtarget.com/opinion/Ten-analytics-success-stories-in-a-nutshell
4. Evans C. Precision medicine is already working to cure Americans: these are their stories; 2015 [Last accessed 17 Dec 2015]. Available from: https://www.whitehouse.gov/blog/2015/01/29/precision-medicine-already-working-cure-americans-these-are-their-stories
5. Highnam G, Mittelman D. Personal genomes and precision medicine. Genome Biol. 2012;13:324.
6. Duffy DJ. Problems, challenges, and promises: perspectives on precision medicine. Brief Bioinform. 2016;1–11. doi:10.1093/bib/bbv060.
7. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13:e1002195.

- **This comparison of genomic data with those from astronomy, social media, and video, now and in 2025, indicates valuable lessons to be learned from other disciplines**.

8. Hu H, Wen Y, Chua T-S, et al. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access. 2014;2:652–687.

•• **This tutorial provides a comprehensive overview of big data generation, acquisition, storage, and analytics technologies across time, domains, and layers of infrastructure, computing, and applications**.

9. Precision Medicine Initiative. The White House; 2015 [Last accessed 17 Dec 2015]. Available from: https://www.whitehouse.gov/precision-medicine
10. Institute of Medicine. Finding what works in health care. National Washington, DC: Academies Press, 2011 [cited 2013 Mar 1]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK209518/.
11. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526:336–342.

•• **The authors review the infrastructure required in the precision medicine ecosystem to integrate material, knowledge and data fully through to clinical interpretation**.

12. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health. 2011;38:65–76.
13. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1:293–314.

- **This overview of common statistical and computational challenges associated with big data analysis provides some suggestions for statistical methods to overcome spurious correlation and bias**.

14. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–94.
15. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14:618–630.
16. Porter JH, Hanson PC, Lin -C-C. Staying afloat in the sensor data deluge. Trends Ecol Evol. 2012;27:121–129.
17. Zhang X, Wu T, Jiang T. A review of EEG and MEG for brainnetome research. Cogn Neurodyn. 2014;8:87–98.
18. Yang J, Gong P, Fu R, et al. The role of satellite remote sensing in climate change studies. Nat Clim Chang. 2013;3:875–883.
19. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
20. Zerhouni E. The NIH roadmap. Science. 2003;302:63–72.
21. Green ED, Guyer MS. National human genome research institute. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470:204–213.
22. Collins FS, Wilder EL, Zerhouni E. NIH Roadmap/Common fund at 10 years. Science. 2014;345:274–276.
23. Eisenstein M. Big data: the power of petabytes. Nature. 2015;527:S2–S4.
24. Sboner A, Elemento O. A primer on precision medicine informatics. Brief Bioinform. 2015. doi:10.1093/bib/bbv032.
25. Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics – the first seven years. Front Genet. 2014;5:184.
26. Verma SS, de Andrade M, Tromp G, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet. 2014;5:370.
27. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.

- **The authors evaluate the outcomes from 5 years of GWAS relative to expectations, and highlight where those expectations may have been unrealistic**.

28. Motsinger-Reif AA, Jorgenson E, Relling MV, et al. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet Genom. 2014;23:383–394.
29. Lippert C, Listgarten J, Liu Y, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–835.
30. Korte A, Vilhjalmsson BJ, Segura V, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44:1066–1071.
31. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–824.
32. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16:321–332.

- **Machine learning approaches have much promise for future mining of big data sources; this overview summarizes current approaches in 'omics and challenges for the future**.

33. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS catalog, a curated resources of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–D1006.
34. EMBL-EBI. GWAS Catalog; 2015 [Last accessed 17 Dec 2015]. Available from: www.ebi.ac.uk/gwas
35. Schrodi SJ, Mukherjee S, Shan Y, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front Genet. 2014;5:162.

•• **The authors review methods for genetic prediction and the gains possible over purely clinical risk prediction with specific examples, concluding that there is still a long way to go for many diseases**.

36. Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, NIH. The Precision Medicine Initiative Cohort Program – building a research foundation for 21st century medicine; 2015 [Last Accessed 17 Dec 2015]. Available from: https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf

•• **This report sketches plans for the NIH-funded Precision Medicine Initiative, which will recruit one million participants starting from 2016 and likely form a basis for similar large-scale studies**.

37. Greely HT. Get ready for the flood of fetal gene screening. Nature. 2011;469:289–291.
38. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450.
39. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–569.
40. Moser G, Lee SH, Hayes BJ, et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11:e1004969.
41. ClinVar. National Center for Biotechnology Information; 2015 [Last accessed 17 Dec 2015]. Available from: https://www.ncbi.nlm.nih.gov/clinvar/
42. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucl Acids Res. 2014;42:D980–D985.
43. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506:185–190.
44. Surakka I, Horikoshi M, Magi R, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015;47:589–597.
45. Mancuso N, Rohland N, Rand KA, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet. 2015. doi:10.1038/ng.3446.

46. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148:1293–1307.
47. Chen R, Snyder M. Systems biology: personalized medicine for the future? Curr Op Pharmac. 2012;12:623–628.
48. Hood L, Lovejoy JC, Price ND. Integrating big data and actionable health coaching to optimize wellness. BMC Medicine. 2015;13:4.

• **The authors argue the case for longitudinal studies integrating genomics, clinical, environmental, and lifestyle measurements, which they claim will ultimately lead to lower costs, better healthcare, innovation, and economic growth**.

49. Diamandis EP. The hundred person wellness project and Google's baseline study: medical revolution or unnecessary and potentially harmful over-testing? BMC Medicine. 2015;13:5.

• **The author argues the case against longitudinal studies of highly quantified individuals, which may not be effective in preventing disease, and may lead to unnecessary and potentially harmful interventions**.

50. Ritchie MD, Holzinger ER, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16:85–97.

•• **This review highlights the challenges of integrating data through meta-dimensional and multi-staged analysis, which will only become more relevant as the magnitude and diversity of data increase**.

51. Sinha R, Abnet CC, White O, et al. The microbiome quality control project: baseline study design and future directions. Genome Biol. 2015;16:276.

• **This recent study identifies issues with variability in microbiome analyses and proposes future studies to evaluate and standardize sample collection, references, techniques, designs, and training**.

52. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. Nat Genet. 2001;29:365–371.
53. Blumenthal D. Stimulating the adoption of health information technology. New Engl J Med. 2009;360:1477–1479.
54. Ben-Assuli O. Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. Health Policy. 2015;119:287–297.

• **This covers the current state of EHR adoption in multiple countries, various obstacles encountered in implementation, and the impact on quality of care**.

55. Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014;5:4022.

• **This large-scale retrospective study of electronic health records from Denmark demonstrates the potential of such longitudinal big data analyses for disease prediction and prevention**.

56. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12:417–418.
57. Flintoft L. Phenome-wide association studies go large. Nat Rev Genet. 2014;15:2.
58. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13:395–405.

•• **This review of data mining for EHR describes the diverse data types encapsulated in the record, the issues inherent in their integration and challenges to using such information for research and clinical care**.

59. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform. 2015;58S:S128–S132.
60. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350:h1885.
61. Ryan PB, Madigan D, Stang PE, et al. Medication-wide association studies. CPT Pharmacometrics Syst Pharmacol. 2013;2:e76.
62. Kukafka R, Ancker JS, Chan C, et al. Redesigning electronic health record systems to support public health. J Biomed Inform. 2007;40:398–409.
63. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc. 2013;20:e147–e154.
64. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances and perspectives. J Am Med Inform Assoc. 2013;20:e206–e211.
65. What is the Phenotype KnowledgeBase? Vanderbilt University; 2014 [Last accessed 17 Dec 2015]. Available from: https://www.phekb.org
66. OHDSI. Observational Health Data Sciences and Informatics; 2015 [Last accessed 17 Dec 2015]. Available from: www.ohdsi.org
67. FDA. Mini-Sentinel; 2014 [Last accessed 17 Dec 2015]. Available from: www.mini-sentinel.org
68. Spiegel PK. The first clinical X-ray made in America – 100 years. Am J Roentgenology. 1995;164:241–243.
69. Arenson RL, Andriole KP, Avrin DE, et al. Computers in imaging and health care: now and in the future. J Digit Imaging. 2000;13:145–156.
70. Seto B, Friedman C. Moving toward multimedia electronic health records: how do we get there? J Am Med Inform Assoc. 2012;19:503–505.
71. Hsu W, Markey MK, Wang MD. Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities. J Am Med Inform Assoc. 2013;20:1010–1013.
72. Public Image Databases. Cornell University: Computer Vision and Image Analysis Group; 2009 [Last accessed 17 Dec 2015]. Available from: http://www.via.cornell.edu/databases/
73. Aylward SR. Open-Access Medical Image Repositories; 2008 [Last accessed 17 Dec 2015]. Available from: http://www.aylward.org/notes/open-access-medical-image-repositories
74. The ADHD-200 Consortium. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Front Syst Neurosci. 2012;6:62.
75. Fonseca CG, Backhaus M, Bluemke DA, et al. The Cardiac Atlas project – an imaging database for computational modeling and statistical atlases of the heart. Bioinformatics. 2011;27:2288–2295.
76. Thompson PM, Stein JL, Medland SE, et al. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 2014;8:153–182.
77. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1046–1057.
78. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19:A68–A77.
79. Clarke LP, Nordstrom RJ, Zhang H, et al. the quantitative imaging network: NCI's historical perspective and planned goals. Transl Oncol. 2014;7:1–4.
80. Smith S. 4 ways the IBM Watson is changing health care, from diagnosing disease to treating it. Medical Daily;2015 [Last accessed 18 Dec 2015]. Available from: http://www.medicaldaily.com/4-ways-ibm-watson-changing-health-care-diagnosing-disease-treating-it-364394
81. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
82. Bron EE, Smits M, Van Der Flier WM, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage. 2015;111:562–579.
83. Lee H, Chen Y-P P. Image based computer aided diagnosis system for cancer detection. Expert Syst Appl. 2015;42:5356–5365.
84. Tokuda O, Harada Y, Ohishi Y, et al. Investigation of computer-aided diagnosis system for bone scans: a retrospective analysis in 406 patients. Ann Nucl Med. 2014;28:329–339.
85. Waterton JC, Pylkkanen L. Qualification of imaging biomarkers for oncology drug development. European J of Cancer. 2012;48:409–415.

86. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. Sci Transl Med. 2015;7:283rv3.
- **This summarizes current and future measurements possible from wearables, obstacles, and incentives for their translation to regular use in health care**.

87. Burke LE, Ma J, Azar KMJ, et al. Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American heart association. Circulation. 2015;132:1157–1213.
88. Donker T, Petrie K, Proudfoot J, et al. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res. 2013;15:e247.
89. Clarke J, Proudfoot J, Birch M-R, et al. Effects of mental health self-efficacy on outcomes of a mobile phone and web intervention for mild-to-moderate depression, anxiety and stress: secondary analysis of a randomised controlled trial. BMC Psychiatry. 2014;14:272.
90. Swan M. The quantified self: fundamental disruption in big data science and biological discovery. Big Data. 2013;1:85–99.
91. Shih PC, Han K, Poole ES, et al. Use and adoption challenges of wearable activity trackers. iConference Proceedings 2015.
92. Patel MS, Asch DA, Volpp KG. Wearable devices as facilitators, not drivers, of health behavior change. JAMA. 2015;313:459–460.
- •• **The authors highlight gaps between recording information and changing behavior for wearables, and conclude that change may rely more on design of engagement strategies than features of the technologies**.

93. Wearing your intelligence: how to apply artificial intelligence in wearables and IoT. WIRED Magazine; 2014 [Last accessed 17 Dec 2015]. Available from: http://www.wired.com/insights/2014/12/wearing-your-intelligence/
94. Banaee H, Ahmed MU, Loutfi A. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors. 2013;13:17472–17500.
95. Brooks C. Introductory econometrics for finance. 3rd ed. Cambridge: Cambridge University Press; 2014.
96. Fan Z, Dror RO, Mildorf TJ, et al. Identifying localized changes in large systems: change-point detection for biomolecular simulations. Proc Natl Acad Sci. 2015;112:1–6.
97. Reis S, Seto E, Northcross A, et al. Integrating modelling and smart sensors for environmental and human health. Environ Modell Software. 2015; 74: 238–246.
- •• **This overview of model-sensor integration identifies key research areas required in human health and environmental analysis to convert big data into 'big information'**.

98. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010;86:560–572.
99. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present and future. Genet Med. 2013;15:761–771.
- **The eMERGE network is a consortium in its third cycle of NHGRI funding, which has taken the lead on integrating genomic and EHR data; this summary of its history and future provides insight for other groups undertaking this integration**.

100. Regeneron and Geisinger Health System announce major human genetics research collaboration. Regeneron Pharmaceuticals, Inc.; 2014 [Last accessed 17 Dec 2015]. Available from: http://investor.regeneron.com/releasedetail.cfm?ReleaseID=818844
101. Kvale MN, Hesselson S, Hoffman TJ, et al. Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genet. 2015;200:1051–1060.
102. Li L, Ruau DJ, Patel CJ, et al., et al. Disease risk factors identified through shared genetic architecture and electronic medical records. Sci Transl Med. 2014;6:234ra57.
103. Li L, Cheng W-Y, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7:311ra174.
104. Shen L, Thompson PM, Potkin SG, et al. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. Brain Imaging Behav. 2014;8:183–207.
105. Stein JL, Hua X, Lee S, et al. Voxelwise genome-wide association study (vGWAS). NeuroImage. 2010;53:1160–1174.
106. Hyde LW, Bogdan R, Hariri AR. Understanding risk for psychopathology through imaging gene-environment interactions. Trends Cogn Sci. 2011;15:417–427.
107. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidem. 2005;14:1847.
108. Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. Environ Mol Mutagen. 2013;54:480–499.
109. Sanchez FM, Gray K, Bellazzi R, et al. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc. 2014;21:386–390.
110. Bradley JR, Holan SH, Wikle CK. Mixed effects modeling for areal data that exhibit multivariate-spatio-temporal dependencies. Arxiv: 1407.7479v2. 2014. Available from: http://arxiv.org/abs/1407.7479
111. Katzfuss M, Cressie N. Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics. 2012;23:94–107.
112. Poldrack RA, Laumann TO, Koyejo O, et al. Long-term neural and physiological phenotyping of a single human. Nat Comm. 2015;6:8885.
113. Mendelsohn J. Personalizing oncology: perspectives and prospects. J Clin Oncol. 2013;31:1904–1911.
114. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J Clin Oncol. 2013;31:1803–1805.
115. Yu Q, Ding J. Precision cancer medicine: where to target? Acta Pharmacol Sin. 2015;36:1161–1162.
116. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417:949–954.
117. Flaherty KY, Puzanov I, Kim KB, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. N Engl J Med. 2010;363:809–819.
118. Lievre A, Bachet J-B, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res. 2006;66:3992–3995.
119. Pao W, Miller V, Zakowski M, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci. 2004;101:13306–13311.
120. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129–2139.
121. Paez JG, Janne PA, Lee JC, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304:1497–1500.
122. Politi K, Herbst RS. Lung cancer in the era of precision medicine. Clin Cancer Res. 2015;21:2213–2220.
123. Fraser M, Berlin A, Bristow RG, et al. Genomic, pathological, and clinical heterogeneity as drivers of personalized medicine in prostate cancer. Urol Oncol-Semin Ori. 2015;33:85–94.
124. Dawson S-J, Rueda OM, Aparicio S, et al. A new genome-driven integrated classification of breast cancer and its implications. Embo J. 2013;32:617–628.
125. Enhance cancer diagnosis & treatment. ASCO CancerLinq; 2015 [Last accessed 17 Dec 2015]. Available from: www.cancerlinq.org
126. Servant N, Romejon J, Gestraud P, et al. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. Front Genet. 2014;5:152.
127. Burrell RA, McGranahan N, Bartek J, et al. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501:338–345.
128. Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene. 2015;34:3215–3225.

129. Jamal-Hanjani M, Hackshaw A, Ngai Y, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. PLoS Biol. 2014;12:e1001906.

• **This exemplar precision medicine prospective study follows lung cancer patients from diagnosis to relapse, aiming to define the spatial and temporal changes in tumor trajectories through sampling and genetic analysis, determining the impact of therapeutic interventions and clonal heterogeneity**.

130. Gajewski TF, Woo S-R, Zha Y, et al. Cancer immunotherapy strategies based on overcoming barriers within the tumor microenvironment. Curr Opin Immun. 2013;25:268–276.
131. Jain RK. Normalizing tumor microenvironment to treat cancer: bench to bedside to biomarkers. J Clin Onc. 2013;31:2205–2218.
132. Andre F, Mardis E, Salm M, et al. Prioritizing targets for precision cancer medicine. Ann Onc. 2014;25:2295–2303.
133. Global Alliance for Genomics & Health. Global Alliance; 2015 [Last accessed 17 Dec 2015]. Available from: https://genomicsandhealth.org/
134. Health Information Privacy. US Department of Health & Human Services; 2015 [Last accessed 17 Dec 2015]. Available from: http://www.hhs.gov/ocr/privacy/index.html
135. Hazin R, Brothers KB, Malin BA, et al. Ethical, legal, and social implications of incorporating genomic information into electronic health records. Genet Med. 2013;15:810–816.
136. Harrison MI, Koppel R, Bar-Lev S. Unintended consequences of information technologies in health care – an interactive sociotechnical analysis. J Am Med Inform Assoc. 2007;14:542–549.
137. Hudson KL, Collins FS. Bringing the common rule into the 21st century. N Engl J Med. 2015;373:2293–2296.
138. Hartzler A, McCarty CA, Rasmussen LV, et al. Stakeholder engagement: a key component of integrating genomic information into electronic health records. Genet Med. 2013;15:792–801.
139. Kohane IS. Ten things we have to do to achieve precision medicine. Science. 2015;349:37–38.

•• **The 10 challenges highlighted here range from technical to sociopolitical and in level of current-day difficulty; on the technical side, computation and dealing with large-scale 'omics are among the most difficult**.

140. Education. American Society for Human Genetics; 2015 [Last accessed 17 Dec 2015]. Available from: http://www.ashg.org/education/
141. Paton C. Massive open online course for health informatics education. Healthc Inform Res. 2014;20:81–87.
142. Brazas MD, Lewitter F, Schneider MV, et al. A quick guide to genomics and bioinformatics training for clinical and public audiences. PLoS Comput Biol. 10: 2014; e1003510.
143. Steinberg D, Horwitz G, Zohar D. Building a business model in digital medicine. Nat Biotech. 2015;33:910–920.

•• **The authors analyze elements of biopharmaceutical, medical device, and technological business models, and propose seven building blocks which will form the basis of new digital medicine models**.

144. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med. 2012;366:489–491.
145. Bourne PE, Lorsch JR, Green ED. Perspective: sustaining the big-data ecosystem. Nature. 2015;527:S16–S17.
146. A community platform for NGS assay evaluation and regulatory science exploration. precisionFDA. U.S. Food and Drug Administration; 2015 [Last accessed 1 Dec 2016]. Available from: https://precision.fda.gov
147. BaseSpace – Genomics Cloud Computing. Illumina, Inc.; 2015 [Last accessed 1 Dec 2016]. Available from: https://basespace.illumina.com/home/prep
148. Marx V. Human phenotyping on a population scale. Nat Methods. 2015;12:711–714.
149. 23andMe Research. 23andMe; 2015 [Last accessed 28 Dec 2015]. Available from: https://www.23andme.com/research/
150. ResearchKit. Apple; 2015 [Last accessed 28 Dec 2015]. Available from: http://www.apple.com/researchkit/?cid=wwa-us-kwg-iphone-com
151. Diehl P. BGI plans to sequence the world; 2013 [Last accessed 28 Dec 2015]. Available from: http://biotech.about.com/od/investinginbiotech/a/Bgi-Plans-To-Sequence-The-World.htm
152. Unrivaled capabilities. deCODE genetics; 2015 [Last accessed 28 Dec 2015]. Available from: http://www.decode.com/research/
153. The 100,000 Genomes Project. Genomics England; 2015 [Last accessed 28 Dec 2015]. Available from: https://medium.com/precision-medicine/the-genome-war-round-two-441c213e542#.9dbpdpu29
154. Robison RJ. The Genome War, round two; 2015 [Last accessed 28 Dec 2015]. Available from: https://medium.com/precision-medicine/the-genome-war-round-two-441c213e542#.9dbpdpu29
155. 100K Wellness Project. Institute of Systems Biology; 2015 [Last accessed 28 Dec 2015]. Available from: https://www.systemsbiology.org/research/100k-wellness-project/
156. Arivale. Arivale; 2015 [Last accessed 1 Dec 2016]. Available from: www.arivale.com
157. Bergen M. Verily, Google's health gambit, is stacked with scientists. Now it needs to build a business. re/code; 2015 [Last accessed 28 Dec 2015]. Available from: http://recode.net/2015/12/14/verily-googles-health-gambit-is-stacked-with-scientists-now-it-needs-to-build-a-business/
158. About BD2K: Data science at NIH. National Institutes of Health; 2015 [Last accessed 17 Dec 2015]. Available from: https://datascience.nih.gov/bd2k/about
159. IMI 2. Innovative Medicines Initiative; 2010 [Last accessed 17 Dec 2015]. Available from: http://www.imi.europa.eu/content/imi-2
160. About ONC. Newsroom, HealthIT.gov; 2014 [Last accessed 17 Dec 2015]. Available from: https://www.healthit.gov/newsroom/about-onc
161. National Health Plan Collaborative. National Health Plan Collaborative; 2014 [Last accessed 17 Dec 2015]. Available from: http://nationalhealthplancollaborative.org/index.html
162. Simons Center for Data Analysis. Simons Foundation; 2015 [Last accessed 17 Dec 2015]. Available from: https://www.simonsfoundation.org/simons-center-for-data-analysis/
163. Welcome. American Society of Clinical Oncology; 2015 [Last accessed 17 Dec 2015]. Available from: http://www.asco.org/genetics-toolkit/welcome
164. Stand up to cancer – this is where the end of cancer begins. Entertainment Industry Foundation; 2015 [Last accessed 17 Dec 2015]. Available from: http://www.standup2cancer.org/
165. Annas GJ, Elias S. 23andMe and the FDA. N Engl J Med. 2014;370:985–988.
166. Baudhuin LM. The FDA and 23andMe: violating the first amendment or protecting the rights of consumers?. Clin Chem. 2014;60:835–837.
167. Williams MS. Is the genomic translational pipeline being disrupted?. Hum Genomics. 2015;9:9.
168. Schork N. Personalized medicine: time for one-person trials. Nature. 2015;520:609–611.
169. The public health evidence for FDA oversight of laboratory developed tests: 20 case studies. Office of Public Health Strategy and Analysis (FDA); [Last accessed 17 Dec 2015]. Available from: http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM472777.pdf
170. McKinsey Global Institute. Big data: the next frontier for innovation, competition, and productivity. June 2011.
171. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotech. 2012;29:613–624.
172. NIMM overview – Health and social care information centre. NHS Infrastructure Maturity Model. Health & Social Care Information Centre; 2015 [Last accessed 21 Dec 2015]. Available from: http://systems.hscic.gov.uk/nimm/overview