International Journal of
Molecular Sciences

MDPI

Review

Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review

Mubashir Hassan [1,2,*], Faryal Mehwish Awan [3], Anam Naz [1], Enrique J. deAndrés-Galiana [4], Oscar Alvarez [5], Ana Cernea [5], Lucas Fernández-Brillet [5], Juan Luis Fernández-Martínez [4] and Andrzej Kloczkowski [2,6,*]

1 Institute of Molecular Biology and Biotechnology (IMBB), The University of Lahore (UOL), Lahore 54590, Pakistan; anam.naz@imbb.uol.edu.pk
2 The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH 43205, USA
3 Department of Medical Lab Technology, The University of Haripur, Haripur 22620, Pakistan; faryal_mehwish@yahoo.com
4 Group of Inverse Problems, Optimization and Machine Learning, University of Oviedo, 33003 Oviedo, Spain; andresenrique@uniovi.es (E.J.d.-G.); jlfm@uniovi.es (J.L.F.-M.)
5 DeepBioInsights, 38311 La Florida, Spain; uo217123@uniovi.es (O.A.); cerneadoina@uniovi.es (A.C.); jlfmuniovi@gmail.com (L.F.-B.)
6 Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH 43205, USA
* Correspondence: mubasher.hassan@nationwidechildrens.org (M.H.); andrzej.kloczkowski@nationwidechildrens.org (A.K.)

Abstract: Big data in health care is a fast-growing field and a new paradigm that is transforming case-based studies to large-scale, data-driven research. As big data is dependent on the advancement of new data standards, technology, and relevant research, the future development of big data applications holds foreseeable promise in the modern day health care revolution. Enormously large, rapidly growing collections of biomedical omics-data (genomics, proteomics, transcriptomics, metabolomics, glycomics, etc.) and clinical data create major challenges and opportunities for their analysis and interpretation and open new computational gateways to address these issues. The design of new robust algorithms that are most suitable to properly analyze this big data by taking into account individual variability in genes has enabled the creation of precision (personalized) medicine. We reviewed and highlighted the significance of big data analytics for personalized medicine and health care by focusing mostly on machine learning perspectives on personalized medicine, genomic data models with respect to personalized medicine, the application of data mining algorithms for personalized medicine as well as the challenges we are facing right now in big data analytics.

Keywords: genomics; big data analytics; personalized medicine; health; computational approaches

Citation: Hassan, M.; Awan, F.M.; Naz, A.; deAndrés-Galiana, E.J.; Alvarez, O.; Cernea, A.; Fernández-Brillet, L.; Fernández-Martínez, J.L.; Kloczkowski, A. Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. *Int. J. Mol. Sci.* **2022**, *23*, 4645. https://doi.org/10.3390/ijms23094645

Academic Editors: Ian A. Nicholls and Vladimir N. Uversky

Received: 28 February 2022
Accepted: 18 April 2022
Published: 22 April 2022

1. Introduction

Personalized medicine is an evolving field of science aimed in using various diagnostic tests to determine which medical treatments will work best for each individual patient. The progress of personalized medicine over the ages can be characterized by several milestones.

1. More than two and a half millennia ago, Hippocrates stated: "every human is distinct, and this affects both the disease prediction and the treatment".
2. In 1956, "favism", the genetic basis for the selective toxicity of fava beans, was discovered to be due to a deficiency in the metabolic enzyme G6PD.
3. In 1985, Renato Dulbecco realized that, in order to advance cancer research, it was necessary to sequence the human genome.
4. In 1988, Genentech Inc. sequenced the entire human growth hormone locus (a world record), making evident the feasibility of sequencing the human genome.
5. In 1990, the Human Genome Project (HGP) was launched, and the first draft was published in 2001, with its final version in 2003.

6. Since the early 1990s, individualized treatments tailored to the genome of each patient have been envisioned but rarely realized.
7. In 1994, a diagnostic test for the prediction of the success of rHGH replacement therapy was developed, being the earliest registry of a companion molecular diagnostics (CMDx) test ever invented.
8. In 1998, when the FDA approved Herceptin (anti-EGFR mAb for EGFR+ breast tumors) and HerceptTest (to detect such tumors), it became the first "official" CMDx invented. Since then, a growing list of diagnostic packages/personalized medicine therapies has received, from the FDA, labels recognizing and recommending them.

The human genome is basically the foundation of personalized medicine, which is considered as the next generation of diagnosis and treatment. This review describes the progress of personalized medicine over time, emphasizing the important milestones achieved through time. Starting from the treatment of malaria, as the first more personalized therapeutic approach, it highlights the need for new diagnostic tools and therapeutic regimens based on the individual's genetic background. Cutting-edge biochemical advances including single-nucleotide polymorphisms (SNPs), genotyping, and biochips have made personalized medicine a reality, justifying the use of the terminology in the last few decades. Variations such as SNPs, insertions and deletions, structural variants, and copy number variations in the human genome play a distinctive role in the manifestation and progression of diseases such as cancer, diabetes, and neurodegenerative and cardiovascular diseases. Hence, biomarkers are being investigated as a way of predicting certain diseases and also to identify patient subgroups that respond only to specific drugs. The discovery of the association between antimalarial drugs and G6PD deficiency has opened up a new perspective regarding the adverse effects of these drugs as well as a more personalized approach to the disease. This was one of the first examples that led to a big step toward the application of a more personalized therapy, which was established as a term many years later in 1991 and is currently still quite limited. Since that time, several clinical trials have proven the efficacy of trastuzumab, also resulting in establishing routine HER-2 testing in breast cancer patients and dramatically changing the therapeutic approach to those carrying the mutation. This gene is a great milestone in applied personalized medicine, clearly showing that the right choice of a drug, based on the genetic background of a patient, can have positive effects on their life.

Massive accumulation of large-scale molecular and clinical data in recent decades has radically changed personalized medicine and has raised great expectations concerning its impact on biomedical research and health care [1,2]. Personalized medicine is a practice of medicine that uses an individual's genetic profile to guide decisions made regarding the prevention, diagnosis, and treatment of disease [3]. Personalized or precision medicine is an emerging medical practice based on a data-driven approach that considers relevant medical, genetic, behavioral, and environmental information about an individual to determine patient-specific therapy [2,4,5]. By linking together diverse datasets to reveal hitherto-unknown casual pathways and correlations, big data allows for far more precision and tailoring than was ever before possible [4]. Recent scientific advancements in high-throughput, high-resolution data-generating technologies enables cost-effective analysis of big datasets on individual health [6]. However, to analyze and integrate such large information, there is a need for new computational approaches such as faster, more integrated processors, larger computer memories, improved sensors, new much sophisticated algorithms, methodologies and cloud computing, which may guide future clinical practice by providing clinically useful information [6,7]. The development of big data approaches has enhanced the ability to probe which parts of biology may have functional and dysfunctional activity. The basic aim of precision medicine is to support the practicing clinician by making that information of pragmatic value. Precision medicine can be succinctly defined as an approach to provide the right treatments to the right patients at the right time [8]. However, for most clinical problems, precision strategies remain aspirational. The challenge of reducing biology to its component parts, then identifying

which can and should be measured to choose an optimal intervention, the patient population that will benefit, and when they will benefit most, cannot be overstated. However, the increasing use of hypothesis-free, big data approaches promises to help us reach this aspirational goal [9].

This review article will offer an overview on recent advancements and an update on important developments in the analysis of big data and future strategies for personalized medicine. Technical and methodological approaches have been systemically discussed elsewhere and we direct the reader to these excellent reviews [10]. Here, we identify key conceptual and infrastructural challenges and provide a perspective on how advances can be and are being used to arrive at precision medicine strategies with specific examples [9].

2. The Conceptualization of Big Data

Distinct dimensions are included in the definition of "big data", namely, volume, velocity, variety, value, variability, visualization, virality, and veracity, which describes the massive volume of structured, semi-structured, and unstructured data (Figure 1) [11–14]. According to the Health Directorate of the Directorate-General for Research and Innovation of the European Commission, big data can be defined as "Big data in health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points" [15]. Various sources of big data in the health care industry and in biomedical research include medical records of patients, results of medical examinations, and hospital records, etc. [16]. In addition, advances in technology have already created and continue to create thousands or even millions of measurements that include the sequencing of DNA, RNA, and the characterization of proteins: their sequence, structure, posttranslational modifications, and function, alongside their clinical features. In order to extract useful information from this huge amount of data, high-end computing solutions, along with appropriate infrastructure to systematically generate and analyze big data, are urgently needed. Moreover, advanced machine learning algorithms and techniques (such as deep learning, and cognitive computing) represent the future toolbox and emerging reality, which can be effectively applied to deliver integrative solutions for multi-view big data analysis in order to explain an event or predict an outcome [2].

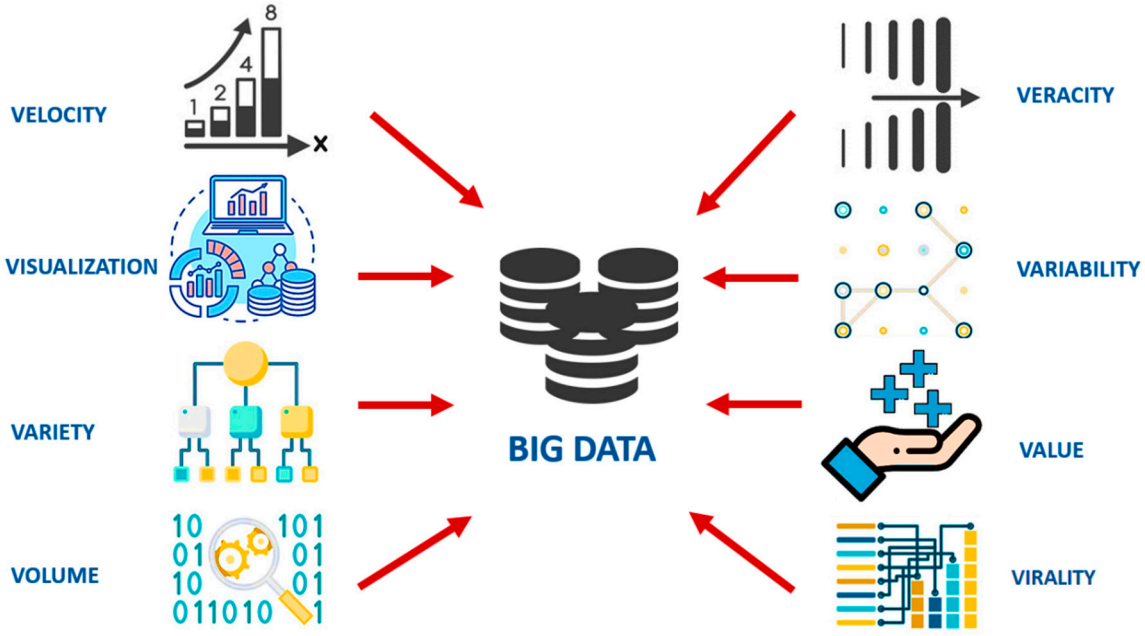

Figure 1. Representation of distinct dimensions of big data.

Despite the recent advancements in machine learning-based solutions for big data, currently, there exist only a few examples that have considerable influence on current clinical practice. Reasons might be a lack of validation via prospective clinical trials, unsatisfactory performance of predictive models, and difficulties in interpreting complex models [16].

It is important to note that when working with genetic data, we should consider that the number of examples (patients) is usually very small in relation to the number of genes or genetic variables that are measured. Therefore, the solution is bounded by the number of patients instead of the number of variables, which makes it *a little big data* problem. This causes the uncertainty space of the mathematical models that are built to solve this kind of problems and make decisions (regressors or classifiers) to have a huge uncertainty space that contains the set of models that predict the observed data within the same error bounds. These models are located in flat curvilinear valleys of the cost function landscape [17,18]. This holds independently of the inverse problem that it is being solved and concerns the uncertainty analysis of inverse problems and classification problems, which are by definition ill-posed. In this way, these problems are very difficult to solve since the noise from the data might dramatically perturb the solution by generating spurious unphysical solutions. Therefore, the best way to deal with such problems is by reducing the dimension to perform a robust uncertainty analysis of the corresponding medical decision problem [19,20]. This kind of approach needs robust sampling methods to consider possible multiple scenarios.

Data formatting and the storing of data also remain as big challenges in the past years. However, the last decade has seen remarkable progress in the development of standard genomic data formats such as FASTQ, BAM/CRAM, and VCF files [21]. However, such standardization is incomplete and may lead to incompatibility between the inputs and outputs of different bioinformatics tools, or worse, inaccurate results. Therefore, imperfect standardization has allowed for the sharing of genomic data across institutions into either aggregated databases such as ExAC, GNOMAD [22] or as federated databases such as the Beacon Network [23]. ExAC, GNOMAD, and the Beacon Network databases provide support in the understanding of genetic variations and identifying variants that are unique within a specific ethnic group [22]. However, despite these successes with upstream genomic data formats, key challenges are still present related to downstream data formats. This often results in non-uniform analysis, and indeed, re-analysis of the same data using different pipelines yields different outcomes [24,25].

3. Computational Approaches toward Personalized Medicine

Personalized medicine refers to the patient's treatment based on their personal clinical characterization [26]. The patient's individual characteristics are used to modify treatment in a way that might be more intricate compared to the standard course [27]. It is evident from recent advances in the pharmacological and genetic behavior of various drugs that genetic variations in a single individual could lead to differences in the response to drugs [28]. All of these factors conspire with the notion of personalized medicine. The main aim of personalized medicine is to achieve the right treatments being given to the right patients.

There has been a rapid development in various high-throughput technologies that has headed toward the addition of a large amount of molecular and cellular biology-related data, providing unprecedented insights into various cellular processes. These computational approaches are now exploiting these extensive data to better understand patient diagnosis, various underlying disease mechanisms, and possible treatment options (Figure 2). Based on genomic, epigenomic profiles, and drug and treatment responses, computational methods can classify the patients into different subtypes that can be helpful in disease prediction, the diagnosis of various cancers, generating disease decision rules, and personalized recommendation systems [29]. These advancements have led many research groups to investigate different aspects of personalized medicine such as diagnosis, prognosis, and pharmacogenomics through computational approaches [30]. Moreover,

such approaches not only refine the existing disease maps but are also beneficial in the development of a predictive model of various diseases [29]. Such analysis is also helpful to differentiate the cellular and molecular mechanism at the normal or control state in comparison to the disease progression state. Thus, these computational approaches for personalized medicine are likely to significantly reshape the therapeutic field in the coming decades. Together, these approaches will allow for the development of various predictive models against various diseases, especially the rare ones.

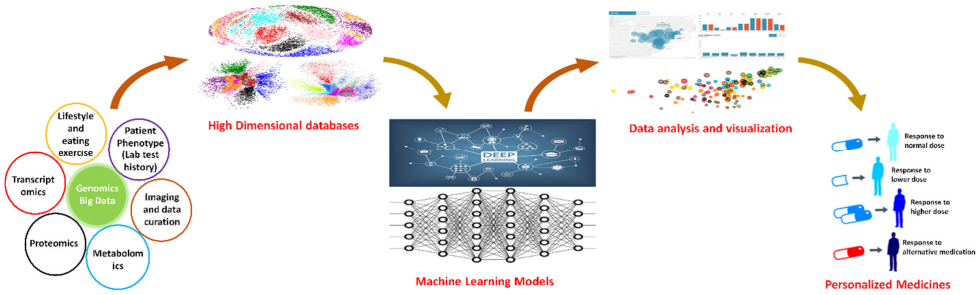

Figure 2. The overall computational approach for personalized medicine.

Nowadays, computational models are integrated in different fields in medicine and drug development, ranging from disease modeling and biomarker research to the assessment of drug efficacy and safety [31]. The added value of such computational models, sometimes called digital evidence, in medicine is also acceptable by the scientific community [32,33] and the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA) [34,35]. There are two types of models: mechanistic models and data derived models. The basic aim of mechanistic models is the structural representation of the governing physiological processes in the model equations to support a functional understanding of the underlying mechanisms. On the other hand, data-driven approaches (machine learning (ML) and deep learning (DL) use algorithms and artificial intelligence (AI) methodology [36,37].

3.1. Molecular Interaction Maps (MIMs)

MIMs actually represent the physical and causal interactions based on knowledge based information among biological species in the form of networks [38]. MIMs explore the information about different mechanistic pathways and regulatory modules involved in a disease such as Parkinson's [39] or signaling in cancer [40], respectively. The basic principle of MIMs uses graph-theory concepts to identify network static properties such as (i) the identification of critical nodes; (ii) community detection; and (iii) prediction of hidden links. Furthermore, upon overlying expression data, such maps serve as visualization tools for the activity level of regulators and their targets of established disease markers, which provide the simplest mechanistic visualization of data [31].

3.2. Constraint-Based Models

Genome-scale metabolic (GEM) models are the best example of constraint-based models that provide a mathematical framework to understand the metabolic capacities of a cell, enabling system wide analysis of genetic perturbations, exploring metabolic diseases, and finding the essential enzymatic reactions and drug targets [41]. Most importantly, the GEM modeling approach is being used in multiple medical domains such as cancer [42] obesity [43], and in Alzheimer's disease [44].

3.3. Boolean Models (BMs)

BMs are the simplest logic-based models in which nodes are assigned one of two possible states: 1 (ON, activation) or 0 (OFF, inactivation) [45]. Moreover, the regulatory relationship between regulators (upstream nodes) to targets (downstream nodes) are expressed by logical operators such as AND, OR, and NOT, respectively. Therefore, BM does

not require detailed kinetic data for parameter estimation, which makes them useful for application to large biological systems. In the context of systems medicine, this approach is often applied for cancer research [46,47].

3.4. Quantitative Models (QMs)

QM are like ordinary differential equation (ODE)-based approaches used to analyze the quantitative behavior of a biochemical reaction with time. QMs consist of a set of differential equations containing variables and parameters that describe how the system responds to different stimuli or perturbations [48]. This quantitative modeling approach explains the biological-systems dynamics in detail and applies to a single pathway due to the requirement of detailed kinetic data for parameter estimations. Most importantly, in personalized medicine, ODE models are applied for individual biomarker discovery [49], drug response, and tailored treatments [50].

3.5. Pharmacokinetic Models

Pharmacokinetic models explain the concentration of a drug in plasma or different tissues. Therefore, drug pharmacokinetics are promptly used as a surrogate for drug-induced responses. Therefore, pharmacokinetic models can be described by compartmental pharmacokinetic (PK) modeling [51] or by physiologically based PK (PBPK) modeling [52].

4. Machine Learning Perspectives on Personalized Medicine

Machine learning imposes a major societal impact in many computational biology applications [53,54]. It has also witnessed dramatic progress as it attempts to identify patterns, rules, and many statistical dependencies in large available datasets. Nowadays, personalized medicine in relation to machine learning programs is considered as an emerging reality and is strongly connected with genomics and proteomics datasets. Machine learning approaches have been applied to massive data collected through genome sequencing, with the aim to precisely define what treatment method will work for an individual [55]. These methodologies have provided deep understanding of the underlying disease mechanisms, while integration of the assorted patient data results in amended and robust biomarker discovery for various disease diagnoses. It has been assessed that without machine learning approaches, the full potential of personalized medicine is impossible to comprehend in clinical practice. Based on machine learning approaches, various algorithms focused on specific diseases have been proposed. Among them, there is an FDA approved MammaPrint prognostic test for breast cancer based on 70 gene signatures [56]. MammaPrint is a microarray-based signature method using formalin-fixed-paraffin-embedded (FFPE) or fresh tissue for microarray analysis [57,58]. Moreover, the BluePrint test has also demonstrated the expression data, which could be supportive for personalized medicine in MINDACT and IMPACt trials [59].

Similarly, Bejnordi et al. reported an algorithm that is trained to detect metastases in various lymph nodes in stained tissue sections of breast cancer [60]. A machine learning echocardiography algorithm proposed by Madani et al. provided an accuracy of greater than 90% for the diagnosis of cardiac disease [61]. For the early detection of Alzheimer's disease, Ding et al. proposed a machine learning based system with high accuracy and sensitivity [27].

Machine learning and AI approaches work with different types of data including genetic, genomic [62], epigenomic [63,64], transcriptomic [65], metabolomic data [66], medical images, biobanks data [67], electronic health records (EHR) [68], scientific literature data, etc., and are able to combine all of this information to design optimum classifiers [69]. In this respect, two problems including regression and classification problems are of interest. The difference between them is that in regression, the aim is to predict the value of continuous and real value quantities, for instance, to predict the level of cholesterol in blood based on other biomarkers. In the case of classification problems, the aim is to predict the label of a set of individuals that are gathered in a broad class, for instance, the

patients that have a survival time greater than the average from the rest. The interest in formulating these prediction problems as classification problems comes from the reduction in the uncertainty space. Particularly, phenotype prediction problems are of great use to better understand the altered genetic pathways that are responsible for the development of the disease and to speed up the drug discovery process [70].

5. Modeling Genetic Data with Translational Purposes

The genetic and epigenetic regulation of the altered pathways in a cell is one of the main topics in pharmacogenomics and consists of understanding how a mutation in the DNA impacts the transcriptome and the proteome downstream [70–72]. Additionally, the epigenomic regulation of the transcriptome can be achieved via epigenomics through chemical compounds that bind to the DNA and alter gene expression [73]. Transcriptomics explores how gene expressions, genetic pathways, and regulatory networks are altered in each phenotype, for instance, disease vs. healthy controls. Based on these findings, it is possible to perform drug repositioning using connectivity map (CMAP) technologies provided by the Broad Institute [74]. Drug repurposing, also called drug repositioning, of the already known FDA approved compounds for new therapeutic uses is a very effective methodology to find a cure in rare diseases where the economic constraints for new drug development are very important [75]. There are multiple examples of personalized medicine being used against multiple diseases. For example, Herceptin (trastuzumab), used in breast cancer, is directed to the 30% of breast cancers with an overexpression of the HER-2 protein, which responds to Herceptin. Gleevec (Imatinib mesylate) is used to treat chronic myeloid leukemia, which has increased life expectancy from 5% to 95% at five years. Zelboraf (Vemurafenib) is used to treat melanoma, where the late-stage prognosis has been dismal, but 60% of patients have a defect in their DNA, and this drug benefits those with the V600E defect. Other successful personalized medicine examples of "treatment–biomarker" combinations are in colon cancer (Erbitux–EFGR) and lung cancer (Xalkori–ALK) [76].

Two other fields of active work are *de novo* drug design [77] and the optimization of gene therapies [78]. Drug discovery is a very challenging problem due to the high attrition rates in *de novo* design due to the lack of the efficacity of the new compounds and due to possible development of undesirable side effects [79–81]. The computational problem consists of finding a new compound that provides the desirable structure–activity relationships (SAR data) [82,83]. This is a very challenging problem due to the high dimensionality of the databases to explore the chemical space, which can be cast as an optimization and/or sampling problem. Local optimization approaches and deep learning methodologies can deal with such problems, but they are unable to perform a complete sampling of the chemical space due to the curse of the dimensionality problem. Additionally, local optimization methods might converge to suboptimal solutions that might be far away from the global solution.

Gene therapy is an experimental technique that uses genes to treat or prevent disease [84]. Several approaches to gene therapy are tested:

1. Replacing a mutated gene that causes disease with a healthy copy of the gene.
2. Inactivating, or "knocking out," a mutated gene that is functioning improperly.
3. Introducing a new gene into the body to help fight a disease.

This promising treatment technique remains risky and requires computational methods to understand the effect of these therapies on gene expression and on proteomics, and how they can affect the health of the patients.

6. Data Mining Tools/Algorithms and Their Applications for Personalized Medicine

The machine learning algorithms are significant to interpret the genomic datasets and help in the design of personalized medicines. The use of multimodal data helps in a deeper analysis of large datasets, which improves the understanding of human health and disease by leaps and bound. Algorithms represent the terminal node in the final predictions from big data [85]. Lee and coauthors proposed a person-centered data mining algo-

rithm that could simultaneously integrate both genetic information and baseline profiles to identify which individual person will benefit from a specific antipsychotic drug among schizophrenic patients. The proposed algorithm can be easily adopted in many other clinical practices for personalized medicine [86]. To analyze metagenomes from novel environmental niches, Ulyantsev and coauthors developed an algorithm named "MetaFast", which enabled them to compare the microflora of a healthy person with the microflora of a patient. As a result, specialists would be able to detect previously unidentified pathogens and their strains, which can aid in the development of personalized medicine [87]. Furthermore, it must be emphasized that the idea behind algorithms is not to replace physicians, but to provide them with tools that support their decisions based on the wealth of available biomedical knowledge and data-driven criteria.

6.1. Pattern-Based Approaches in Data Mining for Analyzing Patient Data

Pattern mining concentrates on identifying rules that describe specific patterns within the data. Pattern mining is the discovery of sequential patterns, for example, sequences of errors or warnings that precede equipment failure may be used to schedule preventative maintenance or may provide insight into a design flaw. Genomic and medical studies have continuously been collecting a huge amount of data on a daily basis and analysis of these data is becoming a challenge with every passing day. Analyzing this huge amount of data needs some practical approaches to deal with it. Sequence data analysis approaches provide several different ways to uncover the precious hidden knowledge in data bulks and to discover novel or important patterns related to a particular disease or individual patient. Here, in this section, we will discuss some of pattern-based approaches (e.g., clustering and temporal pattern analysis) commonly utilized for data mining to analyze the patient's data. Temporal data are more often related to clinical studies and mainly depend on time series, with or without a sequence of events (i.e., the time-based quantitative measurements or sequence of temporal events related to particular clinical study) [88].

6.2. Network Mining for Personalized Medicine and Health Care

To discover the meaningful patterns, interactions, relations, and clinical rules among the variants, data mining and machine learning methods are used to build models for systems biology. Data mining is the "process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis". The medical industry collects a dazzling array of data, most of which are electronic health records (EHRs) collected by HIPAA covered health care facilities. According to a survey by PubMed, data mining is becoming increasingly popular in health care, if not increasingly essential. The huge amounts of data generated by health care EDI transactions cannot be processed and analyzed using traditional methods because of the complexity and volume of the data. Data mining methods include artificial neural networks, clustering, Bayesian networks, decision trees, and genetic algorithms. However, to classify different variants according to the classifications defined by biomedical experts, machine learning techniques are useful, and multiple drug targets could be found by using these techniques. Similar to clustering, the expression of big data at the level of genes and proteins can help in the identification of biomarkers and target candidates [9]. In personalized medicine, medical records representing the very personal biomedical information (individually identifiable health information) are guarded very carefully by the Health Insurance Portability and Accountability Act (HIPAA) and are not available openly. These types of data are not shared centrally to prevent the misuse of big data methods. Furthermore, medical records store the standard medical and clinical data gathered from the patients. There are many errors such as altered data quality and misinterpretations, improper grammatical use, spelling errors, local dialects, and semantic ambiguities, which increase the complexity of data processing and analysis for medical records. Therefore, there is a strong need for the data preprocessing of medical records such as data cleansing, data integration, data transformation, data reduction, and privacy protection [9].

6.3. Big Data Management Problems in Precision Medicine and Health Care

Different types of barriers including philosophical, legal, and practical exist that cause hindrance in the access to data. To improve the translation of big data to health care solutions, several issues need to be addressed that include collecting and standardizing the heterogenous data; data curation; data de-identification and anonymization; legal consents that are required for using the available data; and the importance of providing that data back to the health care communities for further research and usage. As the volumes of big data generated are increasing exponentially, the complexity of these data increases. Sequencing individual human genome is no longer enough, as transcript-level expression analyses of RNA-Seq experiments, metabolites, proteome data, phenotypic and functional traits, can now also be associated within the data. Moreover, earlier research has shown that significant information from a single cell data can provide more details about the biological processes in comparison to the bulk analysis of multiple cell types in a mixed cell population [89]. There are new ways for measuring the single cell genome and transcriptome sequencing (G&T-seq) allowing us to simultaneously obtain both transcriptomic and genomic information from a single cell [90], which can provide clearer information that may be helpful in designing precision medicine [9]. Data volumes are increasing continuously and beyond comprehension, while there is a shortage of bioinformaticians in the current scenario [91]. Prior knowledge about the related domains is considerably helpful to build models based on the big data, so it is suggested, ideally, that analysts should also be trained in biomedicine in addition to bioinformatics. Translational research is usually focused on collecting data abundantly, for example, clinical data, imaging data, genetic and genomic data, and analysts are usually found to be lacking in skills to interpret such types of data [10,62,92].

6.4. Significance of Next Generation Informatics for Big Data in Precision Medicine Era

These approaches can help to transform biomedical data into useful drug development information, and to apply the knowledge for decision support in clinical practice. Today, data integration techniques related to the field of biomedical sciences and health care settings are rapidly revolutionizing research domains by acting as a bridge between biological and medical sciences and data mining. This certainly rests upon a record data upsurge in biological knowledge and research. In the modern era, the field of bioinformatics is facing a challenging task to handle and interpret the massive amounts of genomics, proteomics, and metabolomics data, which are accumulating at an unprecedentedly fast pace. Sparse, noisy, and discontinuous data need special care, which is difficult using traditional machine learning and existing computational methods. Numerous promising solutions have been exploited to tackle big data mining problems and provide creative solutions. In this sense, we must consider that there is no single solution for any biomedical problem. As we have stated in the conceptualization of big data section, there could be infinite models that solve a biomedical problem due to the huge uncertainty space [93]. One of the possible approaches to deal with this ill-posed problem is to sample the uncertainty space using a wide multeity of models. There is no unique model capable of solving this problem perfectly, so we need to explore different techniques, obtaining a solution with its uncertainty assessment, using a consensus strategy [94]. This way, we could give robust information to a medical expert to enhance the medical decision process. There are multiple applications of this methodology in the field of proteomics [95], genomics [96], clinical prognosis [97], cancer treatment [98], aging [99], analysis of defective pathways [100,101], and drug repositioning [102].

The idea is to benefit from biomedical data and apply resourceful informatics approaches to reform the practice of medicine and to improve the health care system. Implementing these approaches promises a bright era of next generation precision medicine [59]. Research strategies that facilitate up-to-date all-encompassing biomedical expertise along with handling vast health care data are highly required.

7. Heterogeneity, a Huge Challenge in Big Data Analysis

Although big data analysis promises great advantages and a potential solution to a diverse range of problems, there remain many unique technical, computational, and statistical challenges that must be addressed to fully explore its potential [103]. Heterogeneity [104], incompleteness, complexity, privacy problems, scalability, lack of structure, storage bottlenecks, spurious correlations, incidental endogeneity, noise accumulation, experimental variations, statistical biases, and measurement errors impede progress at all phases of the big data analysis from data collection and analysis to result in elucidation that can create value from the data [105,106]. In order to solve this issue, data structuring should be the first key step in, or prior, to data analysis. Let us consider a patient with several medical procedures at a hospital, where one record per entire hospital stay or medical procedure or laboratory test can relatively ease the problem of heterogeneity. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Moreover, heterogeneity might be terminological, conceptual, syntactic, or semiotic in nature [104,107]. The problems start right away during data collection as decisions have to be made concerning what data to keep and what to discard, and how to store what is kept reliably. Cloud computing, along with more sophisticated statistical methods, can provide a solution as it offers high scalability, reliability, and autonomy, along with composability and dynamic resource discovery. Moreover, federated learning is also a machine learning method that enables machine learning models to obtain experience from different datasets located at different sites (e.g., local data centers, a central server) without sharing the training data. This allows for personal data to remain in local sites, reducing the possibility of personal data breaches. To handle these challenges, we require transformative solutions, therefore, it is time to develop advanced statistical and computational methods that are robust to data complexity, noises, and data dependence (Figure 3) [103].

Figure 3. Big data challenges in recent times.

8. Role of Big Data in Accelerating Digital Healthcare

Big data analysis will empower digital health care in a way that will ensure the timely access of clinicians to the entire scope of a patient's health information while reducing the need for in-person visits and improving patient outcome [108]. Through digital medicine, big data analysis will prove to be an invaluable tool for health care organizations as it will provide more opportunities for proactive intervention and a more holistic view of the patients' conditions with consolidated real-time information [109]. The health care providers' ability to connect with health care apps to track and monitor patient health will be another key benefit. In addition, via risk modeling and stratification, big data analysis will help to make more accurate predictions of where a patient's health is trending.

By utilizing data-driven insights into patient health, big data analyses will eventually be beneficial for the patients themselves, especially for those who can utilize telemedicine and remote patient monitoring to their advantage, in order to enjoy more flexible and convenient access to care, which in turn will help them to live healthier lives [110–112].

9. Big Data Applications in Health Care

There are so many applications where big data is being implemented to enhance patient care, and, ultimately, can save lives [113]. There are two major divisions of health care big data: vital and social data. The vital category is more significant compared to social big data. However, social big data can also be significant to the health care industry by allowing practitioners to detect attitudes through sentiment analysis [114].

10. Electronic Health Records

Electronic health records (EHRs) are most significant application of big data in medicine and health care [115,116]. The patients must have their medical records reporting demographics, medical history, allergies, and laboratory test results in digital forms. Every record is comprised of one modifiable file, which means that doctors can implement changes over time with no paperwork and no danger of data replication [117].

11. Health Big Data as a Key Player for Informed Strategic Planning

Strategic planning through big data analytics improves the understanding of people's motivations. A common practice is to analyze the check-up results among people in different demographic groups and identify which factors discourage people from taking up treatment. Therefore, better understating of these data and better strategic plans can cure more patients in the most diverse areas [118].

12. Advanced Risk and Disease Management through Big Data

Another significant use of big data is essential for tackling the hospitalization risk for particular patients with chronic diseases [119,120]. More precisely, it can also be used to prevent deterioration. Moreover, by drilling down into insights such as medication type, symptoms, and the frequency of medical visits, among many others, it is possible for health care institutions to provide accurate preventative care and, ultimately, reduce hospital admissions [121].

13. Developing New Therapies and Big Data

Big data in the health care department has the power to discover new medications [122–124]. By using prior data record, real-time, predictive metrics, and cohesive mix data visualization techniques, health care experts can identify the potential strengths and weaknesses in clinical trials or therapeutic processes. Moreover, through data-driven analysis of genetic information as well as efficient predictions for patients, big data analytics in health care can play an important role in the development of new ground breaking drugs and innovative, forward-thinking therapies [125]. Data analytics in health care can streamline, innovate, provide security, save lives, and give confidence and clarity [126].

14. Impediments of Big Data in Health Care

One of the major problems in the use of big data in medicine is that medical data have been collected across different states, hospitals, and administrative departments using different protocols [127–129]. Therefore, new infrastructure resources are required to better cross-examine the medical data through proper collaboration between different data providers. A newly designed software with better efficacy is required for the health care department. We will move from standard regression-based methods, which are a subset of supervised machine learning methods, because the standard regression models can be used

in the machine learning framework to learn from the data and provide outcome predictions based on the inputs [130].

It is known that during online data collection and transmission, health care devices are functional with the help of their MAC addresses and Internet protocol (IP) addresses [97,131]. However, there are serious problem related to the use of this procedure because these network addresses can be accessed and linked to the location and the name of the device's owner, so by analyzing the transmitted data, a hacker could identify an individual including financial and other confidential information. Moreover, there are various open software applications that can track cell phone locations and the names of social media users through MAC and GPS big data, which might be used for malicious reasons [132,133]. Based on such problems, most countries have created legislative principles to secure the personal health care privacy and confidentiality of medical records such as the HIPAA under the Privacy Rule of 2003 in the United States [134,135].

15. Conclusions and Future Prospects

The big data paradigm shift is significantly transforming health care and biomedical research [109,136–139], having the potential to better process clinical and biomolecular information that spans the four dimensions of volume, velocity, variety, and veracity, referring to scale, rate, forms, and content of generated data, respectively [140]. In the current situation, large amount of genomics data could be used to address personalized health care issues [104] and can help to propose new drugs for the treatment of gene related disorders. Advanced machine learning approaches such as artificial intelligence and deep learning represent the future toolbox for the data-driven analytics of genomic big data. The emerging progress in these areas will be indispensable for future innovation in health care and personalized medicine.

Author Contributions: Conceptualization, A.K., M.H. and J.L.F.-M.; writing—original draft preparation, M.H., F.M.A., A.N., E.J.d.-G., O.A., A.C., L.F.-B. and J.L.F.-M.; writing—review and editing, M.H., F.M.A., A.N., E.J.d.-G., O.A., A.C., L.F.-B. and J.L.F.-M.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a NSF grant DBI-1661391, and NIH grants R01GM127701 and R01HG012117. MH acknowledges the Ohio State University for providing the "President's Postdoctoral Scholars Program (PPSP)" award.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Iriart, J.A.B. Precision medicine/personalized medicine: A critical analysis of movements in the transformation of biomedicine in the early 21st century. *Cadernos. Cad. De Saúde Publica* **2019**, *35*. [CrossRef]
2. Cirillo, D.; Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **2019**, *58*, 161–167. [CrossRef] [PubMed]
3. Ginsburg, G.S.; Willard, H.F. Genomic and personalized medicine: Foundations and applications. *Transl. Res.* **2009**, *154*, 277–287. [CrossRef] [PubMed]
4. Schaefer, G.O.; Tai, E.S.; Sun, S. Precision medicine and big data. *Asian Bioeth. Rev.* **2019**, *11*, 275–288. [CrossRef]
5. Naqvi, M.R.; Jaffar, M.A.; Aslam, M.; Shahzad, S.K.; Iqbal, M.W.; Farooq, A. Importance of big data in precision and personalized medicine. In Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 30 July 2020; pp. 1–6.
6. Beckmann, J.S.; Lew, D. Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Med.* **2016**, *8*, 1–11. [CrossRef] [PubMed]
7. Espinal-Enríquez, J.; Mejía-Pedroza, R.; Hernández-Lemus, E. Computational approaches in precision medicine. In *Progress and Challenges in Precision Medicine*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 233–250.
8. Ashley, E.A. Towards precision medicine. *Nat. Rev. Genet.* **2016**, *17*, 507–522. [CrossRef] [PubMed]

9. Hulsen, T.; Jamuar, S.; Moody, A.; Karnes, J.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.; McKinney, E. From big data to precision medicine. *Front. Med.* **2019**, *6*, 34. [CrossRef] [PubMed]
10. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-generation machine learning for biological networks. *Cell* **2018**, *173*, 1581–1592. [CrossRef]
11. Bibault, J.-E. Real-life clinical data mining: Generating hypotheses for evidence-based medicine. *Ann. Transl. Med.* **2020**, *8*, 69. [CrossRef]
12. Normandeau, K. Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. *Inside Big Data* **2013**. Available online: https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ (accessed on 18 January 2022).
13. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
14. Diebold, F.X.; Cheng, X.; Diebold, S.; Foster, D.; Halperin, M.; Lohr, S.; Mashey, J.; Nickolas, T.; Pai, M.; Pospiech, M. A Personal Perspective on the Origin (s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline*. *CiteSeer* **2012**. [CrossRef]
15. Auffray, C.; Balling, R.; Barroso, I.; Bencze, L.; Benson, M.; Bergeron, J.; Bernal-Delgado, E.; Blomberg, N.; Bock, C.; Conesa, A. Making sense of big data in health research: Towards an EU action plan. *Genome Med.* **2016**, *8*, 1–13. [CrossRef] [PubMed]
16. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
17. Fernandez Martinez, J.L.; Fernandez Muniz, M.Z.; Tompkins, M.J. On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics* **2012**, *77*, W1–W15. [CrossRef]
18. Fernández-Martínez, J.L.; Fernández-Muñiz, Z.; Pallero, J.; Pedruelo-González, L.M. From Bayes to Tarantola: New insights to understand uncertainty in inverse problems. *J. Appl. Geophys.* **2013**, *98*, 62–72. [CrossRef]
19. Fernández-Martínez, J.L.; Pallero, J.; Fernández-Muñiz, Z.; Pedruelo-González, L.M. The effect of noise and Tikhonov's regularization in inverse problems. Part I: The linear case. *J. Appl. Geophys.* **2014**, *108*, 176–185. [CrossRef]
20. Fernández-Martínez, J.L.; Pallero, J.; Fernández-Muñiz, Z.; Pedruelo-González, L.M. The effect of noise and Tikhonov's regularization in inverse problems. Part II: The nonlinear case. *J. Appl. Geophys.* **2014**, *108*, 186–193. [CrossRef]
21. Zhang, H. Overview of sequence data formats. In *Statistical Genomics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 3–17.
22. Lek, M.; Karczewski, K.; Minikel, E.; Samocha, K.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.; Ware, J.; Hill, A.; Cummings, B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291. [CrossRef]
23. The Global Alliance for Genomics and Health. A federated ecosystem for sharing genomic, clinical data. *Science* **2016**, *352*, 1278–1280. [CrossRef]
24. Wenger, A.M.; Guturu, H.; Bernstein, J.A.; Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med.* **2017**, *19*, 209–214. [CrossRef] [PubMed]
25. Wright, C.F.; McRae, J.F.; Clayton, S.; Gallone, G.; Aitken, S.; FitzGerald, T.W.; Jones, P.; Prigmore, E.; Rajan, D.; Lord, J. Making new genetic diagnoses with old data: Iterative reanalysis and reporting from genome-wide data in 1133 families with developmental disorders. *Genet. Med.* **2018**, *20*, 1216–1223. [CrossRef] [PubMed]
26. Chan, I.S.; Ginsburg, G.S. Personalized medicine: Progress and promise. *Annu. Rev. Genom. Hum. Genet.* **2011**, *12*, 217–244. [CrossRef] [PubMed]
27. Masoudi-Nejad, A.; Wang, E. Cancer Modeling and Network Biology: Accelerating toward Personalized Medicine. *Semin. Cancer Biol.* **2015**, *30*, 1–3. [CrossRef]
28. Meyer, U.A. Pharmacogenetics–five decades of therapeutic lessons from genetic diversity. *Nat. Rev. Genet.* **2004**, *5*, 669–676. [CrossRef]
29. Janga, S.C.; Edupuganti, M.M.R. Systems and network-based approaches for personalized medicine. *Curr. Synth. Syst. Biol.* **2014**, 2. [CrossRef]
30. Tuena, C.; Semonella, M.; Fernández-Álvarez, J.; Colombo, D.; Cipresso, P. Predictive precision medicine: Towards the computational challenge. In *P5 eHealth: An Agenda for the Health Technologies of the Future*; Springer: Cham, Switzerland, 2020; pp. 71–86.
31. Collin, C.B.; Gebhardt, T.; Golebiewski, M.; Karaderi, T.; Hillemanns, M.; Khan, F.M.; Salehzadeh-Yazdi, A.; Kirschner, M.; Krobitsch, S.; Consortium, E.-S.P. Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation. *J. Pers. Med.* **2022**, *12*, 166. [CrossRef]
32. Apweiler, R.; Beissbarth, T.; Berthold, M.R.; Blüthgen, N.; Burmeister, Y.; Dammann, O.; Deutsch, A.; Feuerhake, F.; Franke, A.; Hasenauer, J. Whither systems medicine? *Exp. Mol. Med.* **2018**, *50*, e453. [CrossRef]
33. Pison, C.; Consortium, C. THE CASyM ROADMAP Implementation of Systems Medicine across Europe; Project Management Jülich, Forschungszentrum Jülich GmbH, Germany. 2014. Available online: https://hal.univ-grenoble-alpes.fr/hal-01969603 (accessed on 28 February 2022).
34. Morrison, T.M.; Pathmanathan, P.; Adwan, M.; Margerrison, E. Advancing regulatory science with computational modeling for medical devices at the FDA's office of science and engineering laboratories. *Front. Med.* **2018**, *5*, 241. [CrossRef]
35. Musuamba, F.T.; Skottheim Rusten, I.; Lesage, R.; Russo, G.; Bursi, R.; Emili, L.; Wangorsch, G.; Manolis, E.; Karlsson, K.E.; Kulesza, A. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. *CPT Pharmacomet. Syst. Pharmacol.* **2021**, *10*, 804–825. [CrossRef]

36. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
37. Dimiduk, D.M.; Holm, E.A.; Niezgoda, S.R. Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. *Integr. Mater. Manuf. Innov.* **2018**, *7*, 157–172. [CrossRef]
38. Kitano, H.; Funahashi, A.; Matsuoka, Y.; Oda, K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol.* **2005**, *23*, 961–966. [CrossRef]
39. Fujita, K.A.; Ostaszewski, M.; Matsuoka, Y.; Ghosh, S.; Glaab, E.; Trefois, C.; Crespo, I.; Perumal, T.M.; Jurkowski, W.; Antony, P. Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.* **2014**, *49*, 88–102. [CrossRef] [PubMed]
40. Kuperstein, I.; Bonnet, E.; Nguyen, H.-A.; Cohen, D.; Viara, E.; Grieco, L.; Fourquet, S.; Calzone, L.; Russo, C.; Kondratova, M. Atlas of Cancer Signalling Network: A systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **2015**, *4*, e160. [CrossRef]
41. Thiele, I.; Palsson, B.Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **2010**, *5*, 93–121. [CrossRef]
42. Uhlen, M.; Zhang, C.; Lee, S.; Sjöstedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F. A pathology atlas of the human cancer transcriptome. *Science* **2017**, *357*, e2507. [CrossRef]
43. Mardinoglu, A.; Agren, R.; Kampf, C.; Asplund, A.; Nookaew, I.; Jacobson, P.; Walley, A.J.; Froguel, P.; Carlsson, L.M.; Uhlen, M. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol. Syst. Biol.* **2013**, *9*, 649. [CrossRef]
44. Stempler, S.; Yizhak, K.; Ruppin, E. Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for Alzheimer's disease. *PLoS ONE* **2014**, *9*, e105383. [CrossRef]
45. Wang, R.-S.; Saadatpour, A.; Albert, R. Boolean modeling in systems biology: An overview of methodology and applications. *Phys. Biol.* **2012**, *9*, 055001. [CrossRef]
46. Eduati, F.; Jaaks, P.; Wappler, J.; Cramer, T.; Merten, C.A.; Garnett, M.J.; Saez-Rodriguez, J. Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Mol. Syst. Biol.* **2020**, *16*, e8664. [CrossRef] [PubMed]
47. Udyavar, A.R.; Wooten, D.J.; Hoeksema, M.; Bansal, M.; Califano, A.; Estrada, L.; Schnell, S.; Irish, J.M.; Massion, P.P.; Quaranta, V. Novel hybrid phenotype revealed in small cell lung cancer by a transcription factor network model that can explain tumor heterogeneity. *Cancer Res.* **2017**, *77*, 1063–1074. [CrossRef] [PubMed]
48. Malik-Sheriff, R.S.; Glont, M.; Nguyen, T.V.; Tiwari, K.; Roberts, M.G.; Xavier, A.; Vu, M.T.; Men, J.; Maire, M.; Kananathan, S. BioModels—15 years of sharing computational models in life science. *Nucleic Acids Res.* **2020**, *48*, D407–D415. [CrossRef] [PubMed]
49. Kolch, W.; Fey, D. Personalized computational models as biomarkers. *J. Pers. Med.* **2017**, *7*, 9. [CrossRef]
50. Hastings, J.F.; O'Donnell, Y.E.; Fey, D.; Croucher, D.R. Applications of personalised signalling network models in precision oncology. *Pharmacol. Ther.* **2020**, *212*, 107555. [CrossRef] [PubMed]
51. Pérez-Urizar, J.; Granados-Soto, V.; Flores-Murrieta, F.J.; Castañeda-Hernández, G. Pharmacokinetic-pharmacodynamic modeling: Why? *Arch. Med. Res.* **2000**, *31*, 539–545. [CrossRef]
52. Edginton, A.N.; Willmann, S. Physiology-based simulations of a pathological condition. *Clin. Pharmacokinet.* **2008**, *47*, 743–752. [CrossRef]
53. Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A. Applications of deep learning in biomedicine. *Mol. Pharm.* **2016**, *13*, 1445–1454. [CrossRef]
54. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
55. Fröhlich, H.; Balling, R.; Beerenwinkel, N.; Kohlbacher, O.; Kumar, S.; Lengauer, T.; Maathuis, M.H.; Moreau, Y.; Murphy, S.A.; Przytycka, T.M. From hype to reality: Data science enabling personalized medicine. *BMC Med.* **2018**, *16*, 1–15. [CrossRef]
56. Cardoso, F.; van't Veer, L.J.; Bogaerts, J.; Slaets, L.; Viale, G.; Delaloge, S.; Pierga, J.-Y.; Brain, E.; Causeret, S.; DeLorenzi, M. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **2016**, *375*, 717–729. [CrossRef] [PubMed]
57. Marchio, C.; Balmativola, D.; Castiglione, R.; Annaratone, L.; Sapino, A. Predictive diagnostic pathology in the target therapy era in breast cancer. *Curr. Drug Targets* **2017**, *18*, 4–12. [CrossRef] [PubMed]
58. Van't Veer, L.J.; Dai, H.; Van De Vijver, M.J.; He, Y.D.; Hart, A.A.; Mao, M.; Peterse, H.L.; Van Der Kooy, K.; Marton, M.J.; Witteveen, A.T. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530–536. [CrossRef] [PubMed]
59. Viale, G.; De Snoo, F.; Slaets, L.; Bogaerts, J.; Van't Veer, L.; Rutgers, E.; Piccart-Gebhart, M.; Stork-Sloots, L.; Glas, A.; Russo, L. Immunohistochemical versus molecular (BluePrint and MammaPrint) subtyping of breast carcinoma. Outcome results from the EORTC 10041/BIG 3-04 MINDACT trial. *Breast Cancer Res. Treat.* **2018**, *167*, 123–131. [CrossRef]
60. Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.; Hermsen, M.; Manson, Q.F.; Balkenhol, M. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **2017**, *318*, 2199–2210. [CrossRef]
61. Madani, A.; Ong, J.R.; Tibrewal, A.; Mofrad, M.R. Deep echocardiography: Data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ Digit. Med.* **2018**, *1*, 1–11. [CrossRef]

62. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef]
63. Rauschert, S.; Raubenheimer, K.; Melton, P.; Huang, R. Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification. *Clin. Epigenet.* **2020**, *12*, 1–11. [CrossRef]
64. Bosco, G.L.; Rizzo, R.; Fiannaca, A.; La Rosa, M.; Urso, A. A deep learning model for epigenomic studies. In Proceedings of the 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, Italy, 28 November–1 December 2016; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA; pp. 688–692.
65. Hamey, F.K.; Göttgens, B. Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. *Exp. Hematol.* **2019**, *78*, 11–20. [CrossRef]
66. Erban, A.; Fehrle, I.; Martinez-Seidel, F.; Brigante, F.; Más, A.L.; Baroni, V.; Wunderlin, D.; Kopka, J. Discovery of food identity markers by metabolomics and machine learning technology. *Sci. Rep.* **2019**, *9*, 9697. [CrossRef]
67. Narita, A.; Ueki, M.; Tamiya, G. Artificial intelligence powered statistical genetics in biobanks. *J. Hum. Genet.* **2021**, *66*, 61–65. [CrossRef] [PubMed]
68. Luz, C.F.; Vollmer, M.; Decruyenaere, J.; Nijsten, M.W.; Glasner, C.; Sinha, B. Machine learning in infection management using routine electronic health records: Tools, techniques, and reporting of future technologies. *Clin. Microbiol. Infect.* **2020**, *26*, 1291–1299. [CrossRef] [PubMed]
69. Lavrač, N. Machine learning for data mining in medicine. In Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, Aalborg, Denmark, 20–24 June 1999; Springer: Aalborg, Denmark; pp. 47–62.
70. Cernea, A.; Fernández-Martínez, J.L.; deAndrés-Galiana, E.J.; Fernández-Ovies, F.J.; Alvarez-Machancoses, O.; Fernández-Muñiz, Z.; Saligan, L.N.; Sonis, S.T. Robust pathway sampling in phenotype prediction. Application to triple negative breast cancer. *BMC Bioinform.* **2020**, *21*, 1–13. [CrossRef] [PubMed]
71. Bonder, M.J.; Kasela, S.; Kals, M.; Tamm, R.; Lokk, K.; Barragan, I.; Buurman, W.A.; Deelen, P.; Greve, J.-W.; Ivanov, M. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genom.* **2014**, *15*, 860. [CrossRef] [PubMed]
72. Niu, N.; Liu, T.; Cairns, J.; Ly, R.C.; Tan, X.; Deng, M.; Fridley, B.L.; Kalari, K.R.; Abo, R.P.; Jenkins, G. Metformin pharmacogenomics: A genome-wide association study to identify genetic and epigenetic biomarkers involved in metformin anticancer response using human lymphoblastoid cell lines. *Hum. Mol. Genet.* **2016**, *25*, 4819–4834. [CrossRef]
73. Liou, S.-Y.; Stringer, F.; Hirayama, M. The impact of pharmacogenomics research on drug development. *Drug Metab. Pharmacokinet.* **2012**, *27*, 2–8. [CrossRef]
74. Lamb, J.; Crawford, E.D.; Peck, D.; Modell, J.W.; Blat, I.C.; Wrobel, M.J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K.N. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929–1935. [CrossRef]
75. Hassan, M.; Raza, H.; Abbasi, M.A.; Moustafa, A.A.; Seo, S.-Y. The exploration of novel Alzheimer's therapeutic agents from the pool of FDA approved medicines using drug repositioning, enzyme inhibition and kinetic mechanism approaches. *Biomed. Pharmacother.* **2019**, *109*, 2513–2526. [CrossRef]
76. Cutter, G.R.; Liu, Y. Personalized medicine: The return of the house call? *Neurol. Clin. Pract.* **2012**, *2*, 343–351. [CrossRef]
77. Hartenfeller, M.; Schneider, G. De novo drug design. *Chem. Inform. Comput. Chem. Biol.* **2010**, 299–323.
78. Álvarez-Machancoses, Ó.; Fernández-Martínez, J.L. Using artificial intelligence methods to speed up drug discovery. *Expert Opin. Drug Discov.* **2019**, *14*, 769–777. [CrossRef]
79. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **2018**, *17*, 97–113. [CrossRef] [PubMed]
80. Gediya, L.K.; Njar, V.C. Promise and challenges in drug discovery and development of hybrid anticancer drugs. *Expert Opin. Drug Discov.* **2009**, *4*, 1099–1111. [CrossRef] [PubMed]
81. Gelb, M.H. Drug discovery for malaria: A very challenging and timely endeavor. *Curr. Opin. Chem. Biol.* **2007**, *11*, 440–445. [CrossRef] [PubMed]
82. Guha, R. On exploring structure–activity relationships. *Silico Models Drug Discov.* **2013**, 81–94.
83. Greene, N.; Fisk, L.; Naven, R.T.; Note, R.R.; Patel, M.L.; Pelletier, D.J. Developing structure− activity relationships for the prediction of hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23*, 1215–1222. [CrossRef]
84. Patil, P.; Chaudhari, P.; Sahu, M.; Duragkar, N. Review article on gene therapy. *Int. J. Genet.* **2012**, *4*, 74.
85. Cahan, E.M.; Hernandez-Boussard, T.; Thadaney-Israni, S.; Rubin, D.L. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit. Med.* **2019**, *2*, 1–6. [CrossRef]
86. Lee, B.S.; McIntyre, R.S.; Gentle, J.E.; Park, N.S.; Chiriboga, D.A.; Lee, Y.; Singh, S.; McPherson, M.A. A computational algorithm for personalized medicine in schizophrenia. *Schizophr. Res.* **2018**, *192*, 131–136. [CrossRef]
87. Ulyantsev, V.I.; Kazakov, S.V.; Dubinkina, V.B.; Tyakht, A.V.; Alexeev, D.G. MetaFast: Fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* **2016**, *32*, 2760–2767. [CrossRef]
88. Bellazzi, R.; Ferrazzi, F.; Sacchi, L. Predictive data mining in clinical medicine: A focus on selected methods and applications. *WIREs Data Min. Knowl. Discov.* **2011**, *1*, 416–430. [CrossRef]
89. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef] [PubMed]

90. Angermueller, C.; Clark, S.J.; Lee, H.J.; Macaulay, I.C.; Teng, M.J.; Hu, T.X.; Krueger, F.; Smallwood, S.A.; Ponting, C.P.; Voet, T. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **2016**, *13*, 229–232. [CrossRef] [PubMed]
91. MacLean, M.; Miles, C. Swift action needed to close the skills gap in bioinformatics. *Nature* **1999**, *401*, 10. [CrossRef]
92. Hood, L. Systems biology and p4 medicine: Past, present, and future. *Rambam Maimonides Med. J.* **2013**, *4*. [CrossRef]
93. Fernández-Martínez, J.L.; Fernández-Muñiz, Z.; Cernea, A.; Pallero, J.; DeAndrés-Galiana, E.J.; Pedruelo-González, L.M.; Álvarez, O.; Fernández-Ovies, F.J. How to deal with uncertainty in inverse and classification problems. In *Advances in Modeling and Interpretation in Near Surface Geophysics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 401–414.
94. deAndrés-Galiana, E.J.; Fernández-Martínez, J.L.; Sonis, S.T. Design of biomedical robots for phenotype prediction problems. *J. Comput. Biol.* **2016**, *23*, 678–692. [CrossRef]
95. Álvarez-Machancoses, Ó.; De Andrés-Galiana, E.J.; Fernández-Martínez, J.L.; Kloczkowski, A. Robust prediction of single and multiple point protein mutations stability changes. *Biomolecules* **2019**, *10*, 67. [CrossRef]
96. deAndrés-Galiana, E.J.; Bea, G.; Fernández-Martínez, J.L.; Saligan, L.N. Analysis of defective pathways and drug repositioning in Multiple Sclerosis via machine learning approaches. *Comput. Biol. Med.* **2019**, *115*, 103492. [CrossRef]
97. deAndrés-Galiana, E.J.; Fernández-Martínez, J.L.; Luaces, O.; del Coz, J.J.; Fernández, R.; Solano, J.; Nogués, E.A.; Zanabilli, Y.; Alonso, J.M.; Payer, A.R. On the prediction of Hodgkin lymphoma treatment response. *Clin. Transl. Oncol.* **2015**, *17*, 612–619. [CrossRef]
98. Reinbolt, R.E.; Sonis, S.; Timmers, C.D.; Fernández-Martínez, J.L.; Cernea, A.; de Andrés-Galiana, E.J.; Hashemi, S.; Miller, K.; Pilarski, R.; Lustberg, M.B. Genomic risk prediction of aromatase inhibitor-related arthralgia in patients with breast cancer using a novel machine-learning algorithm. *Cancer Med.* **2018**, *7*, 240–253. [CrossRef]
99. Cernea, A.; Fernández-Martínez, J.L.; de Andrés-Galiana, E.J.; Fernández-Muñiz, Z.; Bermejo-Millo, J.C.; González-Blanco, L.; Solano, J.J.; Abizanda, P.; Coto-Montes, A.; Caballero, B. Prognostic networks for unraveling the biological mechanisms of Sarcopenia. *Mech. Ageing Dev.* **2019**, *182*, 111129. [CrossRef] [PubMed]
100. Fernández-Martínez, J.L.; de Andrés-Galiana, E.J.; Fernández-Ovies, F.J.; Cernea, A.; Kloczkowski, A. Robust sampling of defective pathways in multiple myeloma. *Int. J. Mol. Sci.* **2019**, *20*, 4681. [CrossRef] [PubMed]
101. deAndrés-Galiana, E.J.; Fernández-Ovies, F.J.; Cernea, A.; Fernández-Martínez, J.L.; Kloczkowski, A. Deep neural networks for phenotype prediction in rare diseases: Inclusion body myositis: A case study. In *Artificial Intelligence in Precision Health*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 189–202.
102. Álvarez-Machancoses, Ó.; deAndrés-Galiana, E.; Fernández-Martínez, J.L.; Kloczkowski, A. In The Utilization of Different Classifiers to Perform Drug Repositioning in Inclusion Body Myositis Supports the Concept of Biological Invariance. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, ICAISC 2020, Zakopane, Poland, 12–14 October 2020; Springer: Berlin/Heidelberg, Germany; pp. 589–598.
103. Fan, J.; Han, F.; Liu, H. Challenges of big data analysis. *Natl. Sci. Rev.* **2014**, *1*, 293–314. [CrossRef]
104. Wang, L. Heterogeneous data and big data analytics. *Autom. Control. Inf. Sci.* **2017**, *3*, 8–15. [CrossRef]
105. Labrinidis, A.; Jagadish, H.V. Challenges and opportunities with big data. *Proc. VLDB Endow.* **2012**, *5*, 2032–2033. [CrossRef]
106. Rahman, J.A. Knowledge Based Trade, Technical Change and Location Environment: The Case of Small and Medium Sized Enterprises Engaged in Advanced Producer Software Services in the South East Region. Ph.D. Thesis, University College London, London, UK, 2005.
107. Alexander, C.A.; Wang, L. Big data analytics in heart attack prediction. *J. Nurs. Care* **2017**, *6*, 2167–2168. [CrossRef]
108. Agrawal, R.; Prabakaran, S. Big data in digital healthcare: Lessons learnt and recommendations for general practice. *Heredity* **2020**, *124*, 525–534. [CrossRef] [PubMed]
109. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef]
110. Cifuentes, C.; Romero, E.; Godoy, J. Design and implementation of a telepediatric primary-level and low-cost system to reduce unnecessary patient transfers. *Telemed. e-Health* **2017**, *23*, 521–526. [CrossRef]
111. Danziger, J.; Ángel Armengol de la Hoz, M.; Li, W.; Komorowski, M.; Deliberato, R.O.; Rush, B.N.; Mukamal, K.J.; Celi, L.; Badawi, O. Temporal trends in critical care outcomes in US minority-serving hospitals. *Am. J. Respir. Crit. Care Med.* **2020**, *201*, 681–687. [CrossRef]
112. Folchetti, L.G.D.; da Silva, I.T.; de Almeida-Pititto, B.; Ferreira, S.R.G. Nutritionists' Health Study cohort: A web-based approach of life events, habits and health outcomes. *BMJ Open* **2016**, *6*, e012081. [CrossRef] [PubMed]
113. Pastorino, R.; De Vito, C.; Migliara, G.; Glocker, K.; Binenbaum, I.; Ricciardi, W.; Boccia, S. Benefits and challenges of Big Data in healthcare: An overview of the European initiatives. *Eur. J. Public Health* **2019**, *29* (Suppl. S3), 23–27. [CrossRef] [PubMed]
114. Zikopoulos, P.; Eaton, C. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*; McGraw-Hill Osborne Media: New York, NY, USA, 2011.
115. Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. *JAMA* **2013**, *309*, 1351–1352. [CrossRef] [PubMed]
116. Wu, P.-Y.; Cheng, C.-W.; Kaddi, C.D.; Venugopalan, J.; Hoffman, R.; Wang, M.D. Omic and electronic health record big data analytics for precision medicine. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 263–273.
117. Khennou, F.; Khamlichi, Y.I.; Chaoui, N.E.H. Improving the use of big data analytics within electronic health records: A case study based OpenEHR. *Procedia Comput. Sci.* **2018**, *127*, 60–68. [CrossRef]

118. Mazzei, M.J.; Noble, D. Big Data and Strategy: Theoretical Foundations and New Opportunities. In *Strategy and Behaviors in the Digital Economy*; IntechOpen: London, UK, 2019.
119. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* **2014**, *33*, 1123–1131. [CrossRef]
120. Dimitrov, D.V. Medical internet of things and big data in healthcare. *Healthc. Inform. Res.* **2016**, *22*, 156–163. [CrossRef]
121. Razzak, M.I.; Imran, M.; Xu, G. Big data analytics for preventive medicine. *Neural Comput. Appl.* **2020**, *32*, 4417–4451. [CrossRef]
122. Leff, D.; Yang, G. Big data for precision medicine. *Engineering* **2015**, *1*, 277–279. [CrossRef]
123. Wooden, B.; Goossens, N.; Hoshida, Y.; Friedman, S.L. Using big data to discover diagnostics and therapeutics for gastrointestinal and liver diseases. *Gastroenterology* **2017**, *152*, 53–67.e3. [CrossRef]
124. Podlesny, N.J.; Kayem, A.V.; Meinel, C. Towards identifying de-anonymisation risks in distributed health data silos. In Proceedings of the International Conference on Database and Expert Systems Applications, Linz, Austria, 26–29 August 2019; Springer: Berlin/Heidelberg, Germany; pp. 33–43.
125. Belle, A.; Thiagarajan, R.; Soroushmehr, S.; Navidi, F.; Beard, D.A.; Najarian, K. Big data analytics in healthcare. *BioMed Res. Int.* **2015**, *2015*, 370194. [CrossRef] [PubMed]
126. Alemayehu, D.; Berger, M.L. Big Data: Transforming drug development and health policy decision making. *Health Serv. Outcomes Res. Methodol.* **2016**, *16*, 92–102. [CrossRef] [PubMed]
127. Wielki, J. Implementation of the big data concept in organizations-possibilities, impediments and challenges. In Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, Kraków, Poland, 8–11 September 2013; pp. 985–989.
128. Furda, R.; Gregus, M. Impediments in healthcare digital transformation. *Int. J. Appl. Res. Public Health Manag. (IJARPHM)* **2019**, *4*, 21–34. [CrossRef]
129. Mathew, P.S.; Pillai, A.S. Big Data solutions in Healthcare: Problems and perspectives. In Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 19–20 March 2015; pp. 1–6.
130. Strang, K.D.; Sun, Z. Hidden big data analytics issues in the healthcare industry. *Health Inform. J.* **2020**, *26*, 981–998. [CrossRef]
131. Wang, H.; Jiang, X.; Kambourakis, G. Special issue on Security, Privacy and Trust in network-based Big Data. *Inf. Sci.—Inform. Comput. Sci. Intell. Syst. Appl. Int. J.* **2015**, *318*, 48–50. [CrossRef]
132. Shen, Y.; Zhang, Y. Transmission protocol for secure big data in two-hop wireless networks with cooperative jamming. *Inf. Sci.* **2014**, *281*, 201–210. [CrossRef]
133. Shull, F. The true cost of mobility? *IEEE Softw.* **2014**, *31*, 5–9. [CrossRef]
134. Brown, B. HIPAA Beyond HIPAA: ONCHIT, ONC, AHIC, HITSP, and CCHIT. *J. Health Care Compliance* **2008**, *10*, 41–44.
135. van Loenen, B.; Kulk, S.; Ploeger, H. Data protection legislation: A very hungry caterpillar: The case of mapping data in the European Union. *Gov. Inf. Q.* **2016**, *33*, 338–345. [CrossRef]
136. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big data: Astronomical or genomical? *PLoS Biol.* **2015**, *13*, e1002195. [CrossRef]
137. Patil, H.K.; Seshadri, R. Big data security and privacy issues in healthcare. In Proceedings of the 2014 IEEE International Congress on Big Data, Washington, DC, USA, 27–30 October 2014; pp. 762–765.
138. Jimenez-Sanchez, G. Genomics innovation: Transforming healthcare, business, and the global economy. *Genome* **2015**, *58*, 511–517. [CrossRef] [PubMed]
139. Martin-Sanchez, F.; Verspoor, K. Big data in medicine is driving big changes. *Yearb. Med. Inform.* **2014**, *23*, 14–20.
140. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Chang.* **2018**, *126*, 3–13. [CrossRef]