HHS Public Access

Author manuscript

Expert Rev Precis Med Drug Dev. Author manuscript; available in PMC 2019 July 08.

Published in final edited form as:

Expert Rev Precis Med Drug Dev. 2019 ; 4(3): 189–200. doi:10.1080/23808993.2019.1617632.

Use of big data in drug development for precision medicine: an update

Tongqi Qian[1], **Shijia Zhu**[2,*], and **Yujin Hoshida**[2,*]

[1]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[2]Liver Tumor Translational Research Program, Simmons Comprehensive Cancer Center, Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Abstract

Introduction: Big-data-driven drug development resources and methodologies have been evolving with ever-expanding data from large-scale biological experiments, clinical trials, and medical records from participants in data collection initiatives. The enrichment of biological- and clinical-context-specific large-scale data has enabled computational inference more relevant to real-world biomedical research, particularly identification of therapeutic targets and drugs for specific diseases and clinical scenarios.

Areas covered: Here we overview recent progresses made in the fields: new big-data-driven approach to therapeutic target discovery, candidate drug prioritization, inference of clinical toxicity, and machine-learning methods in drug discovery.

Expert opinion: In the near future, much larger volumes and complex datasets for precision medicine will be generated, e.g., individual and longitudinal multi-omic, and direct-to-consumer datasets. Closer collaborations between experts with different backgrounds would also be required to better translate analytic results into prognosis and treatment in the clinical practice. Meanwhile, cloud computing with protected patient privacy would become more routine analytic practice to fill the gaps within data integration along with the advent of big-data. To conclude, integration of multitudes of data generated for each individual along with techniques tailored for big-data analytics may eventually enable us to achieve precision medicine.

[*]**Corresponding authors:** Shijia Zhu, Liver Tumor Translational Research Program, Simmons Comprehensive, Cancer Center, Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Y5.222, Dallas, Texas 75390, Shijia.Zhu@UTSouthwestern.edu, Yujin Hoshida, Liver Tumor Translational Research Program, Simmons Comprehensive, Cancer Center, Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Y5.222, Dallas, Texas 75390, Yujin.Hoshida@UTSouthwestern.edu.

Declaration of interest

The authors have no 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

Keywords

Big data; drug development; precision medicine; high-throughput screen; machine learning; in silico drug discovery

1. Introduction

Drug development is a time consuming and complex journey, with a high uncertainty that a drug will actually succeed [1], while the emergence of a big data approach has revolutionized our strategies to tackle long-standing challenges in drug development. The proposal of various high-throughput technologies and collection of multiple-omics data are accelerating the translation of basic research discoveries into clinical practice. The past decade has witnessed the translation of several research results from genomics data to FDA-approval in clinics, and meanwhile, many recently developed data-driven approaches have also shown promising potential for clinical practice. An early and well-know example is aspirin, which is originally used for analgesia treatment, while by integrating the information from the electronic health records (EHRs) of patients, post marketing surveillance data, and pharmacological analysis, researchers also found the potential of aspirin to treat colorectal cancer, and the US Preventive Services Task Force released the draft recommendation on aspirin to prevent colorectal cancer in September 2015 [2]. Similar examples also include raloxifene approved by FDA for invasive breast cancer in 2007, and dapoxetine approved by UK in 2012 for the treatment of premature ejaculation [3]. Big data-driven drug discovery has been an increasingly popular strategy in pharmaceutical companies. Currently, many therapeutics companies have integrated gene-expression analysis, cellular screening systems together with computational healthcare informatics software to identify chemical structures with attributes of interest for oncology drug discovery [4]. Moreover, the high performance combination screening system with cell-based phenotypic assays has been used by pharmaceutical companies to make the combinations of existing compounds to attack multiple dysfunctional pathways in various diseases, e.g., inflammation, respiratory, metabolic and infectious diseases [4]. The aforementioned facts demonstrated that big data has played a significant role in the present drug discovery development. In 2016, we published the paper reviewing the use of big data in drug discovery [5]. As an update, here we will keep discussing the important perspectives in data-driven drug discovery and complement relevant resources useful for broad biomedical research community.

2. New big-data-driven approach to therapeutic target discovery

Big data include diverse arrays of unprecedentedly large datasets [6–16]. Meanwhile, data analysis infrastructure and algorithms tailored to analyze such data have developed rapidly, which significantly facilitate to address big data challenges for paradigm shift towards precision medicine (Table 1). As follows, we will review the therapeutic effect discovery from the perspectives of genomics, proteomics, genome-wide association study (GWAS), pathway, electronic health record (EHR) and phenome-wide association study (pheWAS).

2.1 Genomics and proteomics

Biomarker, as drivers of drug development, have the potential to allow the prediction of health outcomes from perspectives of physical functions and symptoms [17]. The biomarker-based cancer drug development has demonstrated much higher success rate by elevating the rate from a mere 3.4% to the overall oncology drug approval rate 8.3% for 2015 [18]. Genomics has been the forerunner of biomarkers, mainly because of the fact that gene expression profiling is technically easy to obtain. Furthermore, compared to single genomic biomarker, the combinations of genomic biomarkers demonstrated superior prognostic performance [19]. An illustrating example is a combination of 70-gene prognostic signature in sporadic breast tumors, which could allow up to 83% accuracy of poor prognosis [20,21]. In addition to genomics, proteomics also presents the potential to serve as biomarkers. However, due to the challenge remaining in rigorous validation and complex decoding process, proteomics is less likely to be applied for advanced cancer diagnostics to reach required clinical accuracy [22,23]. Consequently, very few multi-parametric tests, until now received 510(k) FDA clearance [24].

2.2 GWAS

An accumulating evidence has demonstrated that genetic targets often suggest effective drug targets [25], where a large amount of supporting examples result from GWA Studies. As a hypothesis-free approach, GWAS can investigate millions of Single nucleotide polymorphisms (SNPs) across the genome at the same time, therefore providing a systematic exploration of the impact of individual genetic variants on a given trait. The GWAS has identified a large number of SNP, trait, common and complex disease relationships [26], which can be used to prioritize genetic findings and further identify therapeutic targets [27–29]. In 2014, a large GWAS conducted by Okada et al. recruiting more than 100,000 rheumatoid arthritis cases and controls, identified 101 SNPs associated with rheumatoid arthritis [30]. Those resulting GWAS SNPs identified targets for 18 out of 27 approved rheumatoid arthritis drugs at that time, and also suggested several novel therapeutics. Similar examples of the power of GWAS were to find drug targets for type 2 diabetes [31] and low density lipoprotein cholesterol levels [32]. An outstanding advantage of GWAS is that it could utilize the hidden power of small effect sizes, particularly, the even nonfunctional association between SNPs and disease, to identify the drug targets with treatable effect on the disease or trait [33].

2.3 Pathway

Beyond GWAS-based single candidate genes, the pathway approach enables finding genes involved in biological pathways or general signaling networks, therefore delivering a list of proteins for drug target discovery [34]. Each pathway includes various interactions among molecules and particular functions are triggered via signaling transmission from molecule to molecule, such as cell division, and therefore, defects in any step of a pathway may cause malfunction of the pathway and furthermore result in disease. The pathway approach is helpful to uncover the underlying relationship between diseases. In the past decades, many researchers failed to identify the possible connections between the apparently different diseases, for instance, osteoporosis and cancer, but the maps of molecular interactions

suggested they could share similarities at a molecular level [35]. As a matter of fact, only a few pathways act crucial roles in disease development, and this principle can be employed to make the connection between diseases, which motivates researchers to pay more attention to those pathways, and explore ways to retain or resume their function to treat a wide range of conditions. With the help of pathway methods, more effective treatments for diseases, such as multiple sclerosis and chronic myelogenous leukemia, have already been delivered. Having a full understanding of pathways and their roles inside the human cells would help break the back of some most difficult-to-treat diseases. Goh et al. originally proposed the strategy to identify the disease–disease relationship based on their shared pathways [36], and furthermore, linked the common disease with the rare disease based on a shared gene [37]. Overall, the pathway/network-based methodologies are of great value in facilitating target identification and subsequent treatment development for diseases [38].

2.4 EHRs

Accumulating evidence has demonstrated the utility of EHR data to identify potential therapeutic drug effects and drug–drug interactions. EHRs may be particularly useful when investigating therapeutic effects, owing to their continuous and longitudinal assessment of clinically relevant outcomes and medication exposures. EHR data has an additional advantage in representing the “real-world” conditions of patient medication use and treatment trends, which is a great resource to uncover the clinical consequences of drug genome interactions. LePendu et al. reproduced known drug adverse events by applying natural language to a large EHR dataset, and further revealed a safety signal of rofecoxib on patients with myocardial infarction, even before this association was identified in a clinical trial [39]. The similar approach also suggested the safety signal of the claudication drug cilostazol on patients with congestive heart failure, in spite of a listed contraindication [40]. An analysis based on 31 EHR-defined drug phenotypes suggested that the use of EHRs could lead to a reduction by 72% in medical expenses, and even decrease the development period. Meanwhile, it also largely elevated the data reuse [41], supported by the evidence that data from more than 90% individuals were used repeatedly [33].

2.5 PheWAS

Integration of EHRs of various disease types from different racial groups to the genomic information delivers a new perspective of precision medicine [42]. Under this circumstance, the phenome-wide association study (PheWAS), which incorporates the information of GWAS and EHRs from large cohort studies, has emerged as another novel and effective paradigm [43]. The PheWAS broadened the scale of genotype–phenotype relationship and enabled researchers to find new uses of old drugs. It is exemplified by the study which integrated the PheWAS and DrugBank [44] to screen possible drug repurposing candidates for both rare and common diseases treatment [45]. A total of 52, 966 drug–disease pairs were discovered in that study, where about one third of these pairs were validated by existing drug disease relationship, ongoing clinical trials and publications, while the remaining could be candidates for future drug repurposing studies [38].

3. Candidate drug prioritization

In terms of the *de novo* drug discovery, various big data resources, such as the chemical structure of small molecules, have been extensively utilized for computational drug discovery. Quantitative structure-activity relationship (QSAR) comprises a series of methods, aiming at modeling the relationship between molecules based on their molecular characteristics, under the assumption that chemicals that fit the same QSAR model may function under the same mechanism [46,47]. Chemical structure-based prioritization of single small molecules and structure family-based pooling of compounds are two traditionally used strategies to computationally infer refined compounds with reduced complexity and cost of drug screening (Table 2). Furthermore, integration of the structure of target protein and biochemical properties of each amino acid residue would enable the better prediction of interaction between small molecules and the targets that they act on.

On the other hand, identification of new therapeutic effects from old drugs leads to the drug repurposing, which has become an alternative to the traditional *de novo* drug discovery approaches, by compensating for their lack of technical efficiency that results in a high failure rate of new approved small-molecular entities [48,49]. Since the basic characteristics of the existing drug are already known such as preclinical, pharmacokinetic, pharmacodynamic and toxicity profiles, the drug repurposing from these drugs can largely shorten the processes of compound development. Accordingly, the compound could step directly into Phase II and III clinical studies, thereby bringing about a lower development cost [49], a high return on investment and an improved development time [50]. As a successful example of drug repositioning, crizotinib, was originally used to treat anaplastic large-cell lymphoma. It has also been developed the new indication for Non-Small Cell Lung Cancer (NSCLC) [51], which outperforms the standard chemotherapy by improving progression-free survival and increasing response rates of NSCLC patients [52]. The drug repurposing, as a promising alternative approach, has been widely utilized for the development of treatments for diseases [38].

Matching signatures by comparing the unique signature of a drug against that of another drug, disease or clinical phenotype, is one of the most widely used drug repurposing approaches to see whether there are similarities suggesting shared biological activity [53,54]. A drug's signature could be obtained from various types of data, that include transcriptomic, proteomic or metabolomic data; chemical structures; or adverse event profiles. Matching transcriptomic signatures is widely used in drug-disease similarity inference [55]. This computational approach is a signature reversion-based strategy by assuming that if a drug can reverse the expression pattern of a hallmark gene sets for a disease of interest, then the drug might provide an effective treatment by reverting the disease phenotype. Although simple, such principles have been successfully applied for treating metabolic diseases [56] and demonstrated great potential to improve novel drug repurposing in a large scale of therapeutic areas [57–59]. The public published transcriptomic data is the main resource for matching transcriptomic signatures. In 2006, the Broad Institute established The Connectivity Map (CMap), which generated transcriptomic profiles by dosing of more than 1,300 compounds in a number of cell lines [60]. The CMap dataset of cellular signatures catalogs transcriptional responses of human cells to chemical

and genetic perturbation, which can be treated as a surrogate phenotypic screen for a large number of compounds and has been successfully exploited to make drug repurposing predictions for a number of disease conditions. The new version of CMap, as part of the NIH LINCS Consortium, is now publicly available at https://clue.io, covering more than a 1,000-fold scale-up of the CMap pilot dataset [61]. This is made possible by a new, low-cost, high-throughput reduced representation expression profiling method called L1000 (978 landmark genes), which can computationally infer the expression levels of 81% of non-measured transcripts. In total, the new CMap has produced 1.3 million L1000 profiles from 27,927 perturbagens (small molecule compounds, shRNAs, cDNAs, and biologics), for a total of 476,251 signatures (consolidating replicates), followed by the computation and annotation tools tailored for their analysis. This enormous resource can also be used along with other public transcriptomic databases (Table 1) and dedicated computational tools [62] to identify novel drug–disease connections and potential drug repositioning opportunities [63,64].

4. Inference of clinical toxicity

It was highlighted by a study that approximately nine of ten drugs, which entered Phase I clinical trials, will eventually failed to be approved by FDA [65]. One of the major reasons is the safety concern, which results in the drug development suspension within 31% of programs [65]. Therefore, there is an urgent requirement of *in silico* approaches to accurately model and efficiently detect the drug toxicity, thereby serving as a higher throughput but less expensive alternative to the conventional drug development processes. The *in silico* method is the essential tool in the early stage of drug discovery to screen low-toxicity compounds, effectively complementing the *in vitro* and *in vivo* toxicity test, and largely improving the overall safety assessment. It also facilitates to minimize the requirement of animal test, therefore, largely reducing the cost and time of toxicity testing.

QSAR model, in addition to drug discovery, can be also used for toxicity prediction through regressions [66,67], using specific toxicity endpoints (e.g., median lethal dose values, tissue-specific toxicity events). For each chemical, the independent variables of regression could be chemical and molecular properties; the dependent variable could be a toxicity endpoint. Through regression, the QSAR model fits a relationship between chemical structures and toxicity that can predict the activities of new chemicals [46,47]. Various QSAR models have been developed, including OECD QSAR toolbox and TopKat (Table 3).

The common challenge for drug toxicity studies was a lack of sufficient and reliable data to learn models. At present, a number of well-defined data from various angles are available online, which greatly facilitate the *in silico* approaches in toxicity prediction [68]. An example of data resource is TOXNET, which incorporated a collection of drug toxicity databases, e.g., HSDB and TOXMAP [69]. ACToR is also a large chemical database that centralizes access to toxicity data [70]. Two important databases incorporated in ACToR are DSSTox [71], which provides a high quality public chemistry resource for improved toxicity prediction and ToxRefDB, which provides chemical toxicity data in a publically accessible format for fast automated screening and assessing chemical exposure, hazard and risk [72]. Other toxicity databases include SuperToxic [73], T3DB [74], and admetSAR [75]. In addition, some bioactivity databases are also public available, for instance, PubChem [76],

ChEMBL [77], and BindingDB [78]. These databases on bioactivity are also important to toxicity prediction [68]. However, a large amount of data has been buried in the archives of the pharmaceutical industry in formats that are hard to utilize [79]. The eTOX project involved collaborations among thirteen pharmaceutical companies, eleven academic institutions, and six enterprises [80], aiming at building a comprehensive toxicity database and providing reliable modeling of drug safety endpoints [79]. These initiatives undertaken would help enable to address current concerns.

Due to the fact that patients might have different genetic backgrounds, the therapeutic window of certain drugs would be also distinct, accordingly raising a more important concern about the personalized drug safety [79]. It could be exemplified by 6-mercaptopurine, which is a drug for acute lymphocytic leukemia and chronic myeloid leukemia. It may take on different side effects in patients with different genetic variants on *TPMT*, *NUDT15*, and *ITPA* [81–83]. Motivated by this fact, genetic tests are necessary to screen patients with specific allele variants beforehand. Theoretically, biomarkers could be learned to predict the drug toxicity for each individual patient, when given enough training datasets. One proof-of-concept example is from the DREAM challenge, in which by integrating genetic profiles of cell lines with the compound chemical information, the *in silico* methods could predict cytotoxicity phenotype [84], largely supporting the feasibility of prediction of individualized drug toxicity [85]. GWAS as a hypothesis-free method has successfully identified novel genes that are responsible for drug response or drug induced toxicity [86]. For example, through GWA study, a genetic variant on the *TCL1A* gene was found to induce musculoskeletal adverse events, revealing the involvement of cytokine receptor genes in the inflammatory response [87]. A similar study found the significant association between the genetic variants on the *CACNB4* gene and the drug-induced alopecia in breast cancer, suggesting the mechanism of the pathogenesis of alopecia involving ion channels [88].

At last, various evidence also suggested the need to look more closely at drug toxicity [89]. Natalizumab was highly effective in treating multiple sclerosis; however, it was associated with the development of progressive multifocal leukoencephalopathy resulting from reactivation of JC virus, therefore leading to its withdrawal [90,91]. In 2006, natalizumab was available again under conditional prescription by investigating seropositivity for JC virus antibodies for patients [90,91]. Such conditional approval of drug prevents its withdrawal, and also reduces the risk of drug toxicity. The similar cases also include re-launches of gefitinib conditional on Epidermal Growth Factor Recepetor activating mutations [92], and perhexiline conditional on poor or intermediate metabolizers of CYP2D6 [93]. These facts all indicated the importance of conditionally re-evaluating drug toxicity and incorporating biomarkers into drug development. This is also consistent with the basic idea of precision medicine.

5. Machine-learning/deep-learning methods in drug discovery

QSAR is a very commonly used technique in the pharmaceutical industry for predicting on-target and off-target activities. QSAR datasets often involve a large number of compounds (>100,000) and descriptors (>1,000), and therefore, prioritizing drug compounds from

QSAR is often computationally intensive and requires the adjustment of many sensitive parameters to achieve good prediction [94]. To address these challenges, various machine learning methods have been applied to QSAR, such as linear discriminant analysis [95], k nearest neighbors [96], decision tree [97], support vector machine [98] and random forest [99]. In particular, random forest has been very popular since it was introduced as a QSAR method [99]. Random forests, owing to its high performance, ease of use, and robustness to adjustable parameters, have been treated as “gold standard”, so that other QSAR methods often compared to Random forest to justify their own accuracies [100]. Random forests are an ensemble learning method, which correct for decision trees’ habit of over-fitting to their training set by building multiple decision trees and merging them together to get a more accurate and stable prediction. Random forests have been widely used for bioactivity classification [101], toxicity modeling [102], and drug target identification [103], and so on.

Although promising, the above methods are still considered to be shallow in terms of learning capability as compared to deep learning [104]. The concept of deep learning was built on the basis of artificial neural networks, where the feedforward neural networks combined with many hidden layers is considered to be the deep neural network [105]. Many hidden layers are incorporated so that more abstract patterns can be recognized from input data, where the lower layer can learn basic patterns and upper layers learn higher-level patterns. The well-developed deep learning algorithms include the convolutional neural network (CNN), deep belief networks (DBN), recurrent neural network (RNN), and deep auto-encoder networks (DAENs). The deep learning, although complex, is still composed of many simple and nonlinear processing units, which can extract the features from lower levels, and convert them into new forms of features at a higher, more abstract and more representative level [106]. The aforementioned traditional algorithms have difficulty in dealing with data in the raw form; therefore, it is crucial for them to extract features manually to represent the data, however, this is often intractable, and requires expertise in the specific area of input data. Deep learning algorithms, as opposed to the traditional algorithms, have the capability of automatically extracting useful features directly from the raw data, which are used as new representations facilitating the better performance for further classification [104]. Therefore, such new theory has caught the attention of many researchers and pharmaceutical companies. However, the limitation by deep learning models is that, in spite of high prediction performance, deep learning methods still have difficulties in revealing and interpreting the associated biological mechanisms directly from the data by working as ‘black boxes’ [107].

Drug repurposing is an effective strategy to find new indications from existing drugs. The methods for QSAR are potentially useful for drug repurposing, but some methods tailored for drug repurposing are also proposed. Network-based cluster approaches are a widely used method for drug repurposing. Motivated by the fact that biologic entities of both disease and drug, in the same module of biological networks share similar properties, the network-based clustering approaches as well as network-based propagation approaches have been proposed, aiming at discovering modules (subnetworks, groups or cliques), which can reveal various novel relationships such as drug-disease, drug-drug or drug-target relationships [108]. The most widely used network-based cluster methods include DBSCAN [109], CLIQUE [110], STING [111], and OPTICS [112]. In addition to the network-based approaches, the semantic

and text-mining approaches screen hundreds of thousands of published literatures, enabling the extraction of various biological concepts. Semantic knowledge graphs are constructed to connect biological entities utilizing literature knowledge and biological databases, and furthermore, knowledge graphs can be used to infer novel connections based on network mining methods [113,114]. Semantics-based approaches take full advantage of semantics information included in massive amounts of literatures, thereby, further improving the prediction accuracy of biological entity relationships [108]. Chen et al. [113] integrates multiple datasets, including drugs, targets, and disease pathways, reconstructing a huge semantic network among over 290,000 nodes, followed by a statistical model to predict drug-target relationships. For instance, barbiturate, a drug used for treating migraines, was predicted for use in curing insomnia, which is also supported by literatures [108].

As available big datasets and public archives have made major strides in the past decade. Study shows that every 18 months, the size of public raw sequencing data is doubled. The International Cancer Genome Consortium (ICGC) [115], collecting samples from more than 20,000 donors, aligned sequencing reads for over 1 PB of storage. Moreover, such large ongoing studies as Precision Medicine Initiative [116] and Million Veterans Project [117] will keep raising the total amount even more rapidly. Utilizing these data for drug discovery depends on the large-scale computational resources. Cloud computing, which is well-known for its elasticity, reproducibility and confidentiality features, can ideally serve the large-scale reanalysis of both public and privacy data, and is widely used for research and large-scale collaborations [118], by providing a platform where users rent computers and storage from large data centers. Currently, multiple cloud service platforms are commercially available (Table 4). It is a timely solution to address the huge required computational resources posed by the big data driven genomics research [118].

6. Expert opinion

Drug development is costly and slow, with medications failing due to lack of efficacy or presence of toxicity. The methodologies tailored for big data integrate information from various perspectives, facilitate screening drug candidates, and enable drug toxicity prediction, thereby largely accelerating the drug development and improving medication efficacy. The precision medicine aims at prevention and treatment strategies with individual heterogeneity taken into account. Its dependence on big data will increase along with the personalized medicine gradually coming to the fore. In the near future, much larger volumes and complex datasets for precision medicine will be generated. Large cohorts will be recruited across countries, e.g., the pledge of sharing one million genomes across thirteen European countries [119] and Human Longevity Institute [120], with individual multi-omics, EHRs and environmental factors recorded longitudinally, largely facilitating the study of both genetic heterogeneities within populations and gene-environment interactions. Meanwhile, it also raises promising potential for interrogating even more difficult diseases, such as rare diseases. As a huge potential market, there are more than 6000 rare diseases that has been reported, while 95% of them remain to be studied [121]. Furthermore, with larger data generated, more and even closer collaborations between experts with different background, including computational scientists, cell and molecular biologists, and clinical doctors will be required to better facilitate the translation from analytic results to prognosis

and treatment in the clinical practice. With cloud computing becoming more routine analytic practice, more sophisticated methodologies, will be advanced to fill the gaps within data integration and standardization across different individuals, samples, and modalities along with the advent of big-data. Although promising, the cloud computing also raises higher security concern of patient privacy. The potential solution could be enhancing data governing via private cloud, which saves the critical data in-house, but increases the expense for data analytics. The balance should be considered carefully between the budget of cloud choices and patient privacy protection when adopting big data analytics [122]. Moreover, with the direct-to-consumer (DTC) genomic sequencing being more affordable, genomic analysis would become part of the routine clinical care, and consequently, more individual sequencing data would be generated for patients. The genetic risks for a wide range of medical conditions and common diseases (e.g., heart disease, diabetes) could be learned. This largely facilitate the personalized treatment along with proven options for screening and risk reduction via health behaviors. However, care is also needed, since there are still concerns over the DTC data, for instance, discordance of results between companies, variable quality of pre-test and post-test information, and the lack of medical supervision among DTC companies [89]. On the other hand, the generation of large personal data also poses the new challenges for data sharing. Policy statement would include more flexibility in order to maximize information sharing and increase of data utility. The tools, such as genomic data commons [123], which can accept donations of genomic information from willing patients, would become much more important to enable such efforts. Furthermore, the dedicated characteristics in precision medicine make even harder the preclinical testing of drug and translational medical practice. To fill the gaps, the organs-on-a-chip [124] and personalized stem cell therapies [125] will deliver important contribution to development in precision medicine. Since Artificial intelligence systems of future will become better over time, it would consider many variables and assign different factors in order of their importance, and it would come up with multiple possible candidate diagnoses within a short period of time, with associated probabilities that might improve accuracy and efficiency, and this might change the current trend that physician time with patients is limited. In the future, physicians might spend more time with patients [126].

Funding

This work is supported by U.S. NIH/NIDDK R01 DK099558, European Union ERCE2014EAdGE671231

HEPCIR, Irma T. Hirschl Trust, U.S. Department of Defense W81XWHE16E1E0363, and Cancer Prevention and Research Institute of Texas RR180016 (to Y Hoshida).

References

Articles of special interest have been highlighted as either of interest (*) or of considerable interest (**) to readers.

1. Altevogt BM, Davis M, Pankevich DE, Norris SMP. Improving and Accelerating Therapeutic Development for Nervous System Disorders: Workshop Summary (National Academies Press, 2014).
2. Force UPST. Aspirin use to prevent cardiovascular disease and colorectal cancer: preventive medication. (Ed.^(Eds) (2016)

3. Pushpakom S, Iorio F, Eyers PA et al. Drug repurposing: progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41 (2019). [PubMed: 30310233]
4. Cha Y, Erez T, Reynolds I et al. Drug repurposing from the perspective of pharmaceutical companies. British journal of pharmacology, 175(2), 168–180 (2018). [PubMed: 28369768]
5. Kim RS, Goossens N, Hoshida Y. Use of big data in drug development for precision medicine. Expert review of precision medicine and drug development, 1(3), 245–253 (2016). [PubMed: 27430024]
6. Dimitrakopoulou K, Dimitrakopoulos GN, Sgarbas KN, Bezerianos A. Tamoxifen integromics and personalized medicine: dynamic modular transformations underpinning response to tamoxifen in breast cancer treatment. Omics: a journal of integrative biology, 18(1), 15–33 (2014). [PubMed: 24299457]
7. Jain A, Rakhi N, Bagler G. Analysis of food pairing in regional cuisines of India. PloS one, 10(10), e0139539 (2015). [PubMed: 26430895]
8. Higdon R, Earl RK, Stanberry L et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. Omics: a journal of integrative biology, 19(4), 197–208 (2015). [PubMed: 25831060]
9. Özdemir V, Faris J, Srivastava S. Crowdfunding 2.0: the next-generation philanthropy: A new approach for philanthropists and citizens to co-fund disruptive innovation in global health. EMBO reports, 16(3), 267–271 (2015). [PubMed: 25656538]
10. Calimlioglu B, Karagoz K, Sevimoglu T, Kilic E, Gov E, Arga KY. Tissue-specific molecular biomarker signatures of type 2 diabetes: an integrative analysis of transcriptomics and protein–protein interaction data. Omics: a journal of integrative biology, 19(9), 563–573 (2015). [PubMed: 26348713]
11. Network CGAR. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. New England Journal of Medicine, 368(22), 2059–2074 (2013). [PubMed: 23634996]
12. Network CGAR. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. New England Journal of Medicine, 372(26), 2481–2498 (2015). [PubMed: 26061751]
13. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57 (2012). [PubMed: 22955616]
14. Consortium EP. The ENCODE (ENCyclopedia of DNA elements) project. Science, 306(5696), 636–640 (2004). [PubMed: 15499007]
15. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature, 489(7414), 109 (2012). [PubMed: 22955621]
16. Djebali S, Davis CA, Merkel A et al. Landscape of transcription in human cells. Nature, 489(7414), 101 (2012). [PubMed: 22955620]
17. Liu CH, Abrams ND, Carrick DM et al. Biomarkers of chronic inflammation in disease development and prevention: challenges and opportunities. (Ed.^(Eds) (Nature Publishing Group, 2017)
18. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. (Ed.^(Eds) (Massachusetts Institute of Technology, Department of Electrical Engineering …, 2017)
19. Borrebaeck CAJNRC. Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer. 17(3), 199 (2017). [PubMed: 28154374]
20. Van·t Veer LJ, Dai H, Van De Vijver MJ et al. Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871), 530 (2002). [PubMed: 11823860]
21. Van De Vijver MJ, He YD, Van·t Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine, 347(25), 1999–2009 (2002). [PubMed: 12490681]
22. Schully SD, Carrick DM, Mechanic LE et al. Leveraging biospecimen resources for discovery or validation of markers for early cancer detection. JNCI: Journal of the National Cancer Institute, 107(4) (2015).
23. Ransohoff DF. Proteomics Research to Discover Markers: What Can We Learn from Netflix®? Clinical chemistry, 56(2), 172–176 (2010). [PubMed: 20040622]

24. Füzéry AK, Levin J, Chan MM, Chan DW. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. Clinical proteomics, 10(1), 13 (2013). [PubMed: 24088261]
25. Sanseau P, Agarwal P, Barnes MR et al. Use of genome-wide association studies for drug repositioning. Nature biotechnology, 30(4), 317 (2012).
26. Hindorff LA, Sethupathy P, Junkins HA et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences, 106(23), 9362–9367 (2009).
27. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nature reviews Drug discovery, 12(8), 581 (2013). [PubMed: 23868113]
28. Wang Z-Y, Zhang H-Y. Rational drug repositioning by medical genetics. Nature biotechnology, 31(12), 1080 (2013).
29. Nelson MR, Tipney H, Painter JL et al. The support of human genetic evidence for approved drug indications. Nature genetics, 47(8), 856 (2015). [PubMed: 26121088]
30. Okada Y, Wu D, Trynka G et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 506(7488), 376 (2014). [PubMed: 24390342]
31. Florez JC. Mining the genome for therapeutic targets. Diabetes, dbi160069 (2017).
32. Kathiresan S, Melander O, Guiducci C et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature genetics, 40(2), 189 (2008). [PubMed: 18193044]
33. Denny JC, Driest SL, Wei WQ, Roden DM. The influence of big (clinical) data and genomics on precision medicine and drug development. Clinical Pharmacology & Therapeutics, 103(3), 409–418 (2018). [PubMed: 29171014]
34. Jin G, Wong ST. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug discovery today, 19(5), 637–644 (2014). [PubMed: 24239728]
35. Disease pathways: A key to new drug discovery. (Ed.^(Eds) (NOVARTIS.com, 2013)
36. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proceedings of the National Academy of Sciences, 104(21), 8685–8690 (2007).
37. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PloS one, 4(2), e4346 (2009). [PubMed: 19194489]
38. Delavan B, Roberts R, Huang R, Bao W, Tong W, Liu Z. Computational drug repositioning for rare diseases in the era of precision medicine. Drug discovery today, 23(2), 382–394 (2018). [PubMed: 29055182]
39. LePendu P, Iyer SV, Bauer-Mehren A et al. Pharmacovigilance using clinical notes. Clinical pharmacology & therapeutics, 93(6), 547–555 (2013). [PubMed: 23571773]
40. Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PloS one, 8(5), e63499 (2013). [PubMed: 23717437]
41. Bowton E, Field JR, Wang S et al. Biobanks and electronic medical records: enabling cost-effective research. Science translational medicine, 6(234), 234cm233–234cm233 (2014).
42. Roden DM, Denny JC. Integrating electronic health record genotype and phenotype datasets to transform patient care. Clinical Pharmacology & Therapeutics, 99(3), 298–305 (2016). [PubMed: 26667791]
43. Denny JC, Bastarache L, Ritchie MD et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology, 31(12), 1102 (2013).
44. Law V, Knox C, Djoumbou Y et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research, 42(D1), D1091–D1097 (2013). [PubMed: 24203711]
45. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nature biotechnology, 33(4), 342 (2015).
46. Devillers J Methods for building QSARs In: Computational Toxicology. (Springer, 2013) 3–27.
47. Roy K, Kar S, Das RN. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment (Academic press, 2015).

48. Munos BH, Chin WW. How to revive breakthrough innovation in the pharmaceutical industry. Science translational medicine, 3(89), 89cm16–89cm16 (2011).
49. Mignani S, Huber S, Tomas H, Rodrigues J, Majoral J-P. Why and how have drug discovery strategies in pharma changed? What are the new mindsets? Drug discovery today, 21(2), 239–249 (2016). [PubMed: 26376356]
50. Challenges Novac N. and opportunities of drug repositioning. Trends in pharmacological sciences, 34(5), 267–272 (2013). [PubMed: 23582281]
51. Deotarse PP JAS, Baile MB, Kolhe NS, Kulkarni AA Drug Repositioning: A Review International Journal of Pharma Research & Review, (2015).
52. Shaw AT, Kim D-W, Nakagawa K et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. New England Journal of Medicine, 368(25), 2385–2394 (2013). [PubMed: 23724913]
53. Keiser MJ, Setola V, Irwin JJ et al. Predicting new molecular targets for known drugs. Nature, 462(7270), 175 (2009). [PubMed: 19881490]
54. Hieronymus H, Lamb J, Ross KN et al. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer cell, 10(4), 321–330 (2006). [PubMed: 17010675]
55. Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Briefings in bioinformatics, 12(4), 303–311 (2011). [PubMed: 21690101]
56. Wagner A, Cohen N, Kelder T et al. Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia. Molecular systems biology, 11(3), 791 (2015). [PubMed: 26148350]
57. Hsieh Y-Y, Chou C, Lo H, Yang P. Repositioning of a cyclin-dependent kinase inhibitor GW8510 as a ribonucleotide reductase M2 inhibitor to treat human colorectal cancer. Cell death discovery, 2, 16027 (2016). [PubMed: 27551518]
58. Kunkel SD, Suneja M, Ebert SM et al. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. Cell metabolism, 13(6), 627–638 (2011). [PubMed: 21641545]
59. Malcomson B, Wilson H, Veglia E et al. Connectivity mapping (ssCMap) to predict A20-inducing drugs and their antiinflammatory action in cystic fibrosis. Proceedings of the National Academy of Sciences, 113(26), E3725–E3734 (2016).
60. Lamb J, Crawford ED, Peck D et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science, 313(5795), 1929–1935 (2006). [PubMed: 17008526]
61. Subramanian A, Narayan R, Corsello SM et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 171(6), 1437–1452. e1417 (2017). [PubMed: 29195078]
62. Wang Z, Monteiro CD, Jagodnik KM et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nature communications, 7, 12846 (2016).
63. Pacini C, Iorio F, Gonçalves E et al. DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. Bioinformatics, 29(1), 132–134 (2012). [PubMed: 23129297]
64. Zhang S-D, Gant TW. sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. BMC bioinformatics, 10(1), 236 (2009). [PubMed: 19646231]
65. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature biotechnology, 32(1), 40 (2014).
66. Patlewicz G, Fitzpatrick JM. Current and future perspectives on the development, evaluation, and application of in silico approaches for predicting toxicity. Chemical research in toxicology, 29(4), 438–451 (2016). [PubMed: 26686752]
67. Raies AB, Bajic VB. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdisciplinary Reviews: Computational Molecular Science, 6(2), 147–172 (2016). [PubMed: 27066112]
68. Yang H, Sun L, Li W, Liu G, Tang Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. Frontiers in chemistry, 6, 30 (2018). [PubMed: 29515993]

69. Fowler S, Schnall JG. TOXNET: information on toxicology and environmental health. AJN The American Journal of Nursing, 114(2), 61–63 (2014).
70. Judson R, Richard A, Dix D et al. ACToR—aggregated computational toxicology resource. Toxicology and applied pharmacology, 233(1), 7–13 (2008). [PubMed: 18671997]
71. Williams-DeVane CR, Wolf MA, Richard AM. DSSTox chemical-index files for exposure-related experiments in ArrayExpress and Gene Expression Omnibus: enabling toxico-chemogenomics data linkages. Bioinformatics, 25(5), 692–694 (2009). [PubMed: 19158160]
72. Martin M, Judson R. ToxRefDB-Release user-friendly web-based tool for mining ToxRefDB. Washington, DC: US Environmental Protection Agency,(2010).
73. Schmidt U, Struck S, Gruening B et al. SuperToxic: a comprehensive database of toxic compounds. Nucleic acids research, 37(suppl_1), D295–D299 (2008). [PubMed: 19004875]
74. Wishart D, Arndt D, Pon A et al. T3DB: the toxic exposome database. Nucleic acids research, 43(D1), D928–D934 (2014). [PubMed: 25378312]
75. Cheng F, Li W, Zhou Y et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. (Ed.^(Eds) (ACS Publications, 2012)
76. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic acids research, 37(suppl_2), W623–W633 (2009). [PubMed: 19498078]
77. Gaulton A, Hersey A, Nowotka M et al. The ChEMBL database in 2017. Nucleic acids research, 45(D1), D945–D954 (2016). [PubMed: 27899562]
78. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research, 44(D1), D1045–D1053 (2015). [PubMed: 26481362]
79. Jiang P, Sellers WR, Liu XS. Big Data Approaches for Modeling Response and Resistance to Cancer Drugs. Annual Review of Biomedical Data Science, 1, 1–27 (2018).**This paper summarizes the recent advancement in big data methods for cancer drug efficacy, and discusses the potential opportunities and challenges.
80. Sanz F, Pognan F, Steger-Hartmann T et al. Legacy data sharing to improve drug safety assessment: the eTOX project. Nature Reviews Drug Discovery, 16(12), 811 (2017).
81. Yang JJ, Landier W, Yang W et al. Inherited NUDT15 variant is a genetic determinant of mercaptopurine intolerance in children with acute lymphoblastic leukemia. Journal of clinical oncology, 33(11), 1235 (2015). [PubMed: 25624441]
82. Pratt V, McLeod H, Dean L, Malheiro A, Rubinstein W. Mercaptopurine Therapy and TPMT Genotype--Medical Genetics Summaries. (2012).
83. de Beaumais TA, Fakhoury M, Medard Y et al. Determinants of mercaptopurine toxicity in paediatric acute lymphoblastic leukemia maintenance therapy. British journal of clinical pharmacology, 71(4), 575–584 (2011). [PubMed: 21395650]
84. Eduati F, Mangravite LM, Wang T et al. Prediction of human population responses to toxic compounds by a collaborative competition. Nature biotechnology, 33(9), 933 (2015).
85. Abdo N, Xia M, Brown CC et al. Population-based in vitro hazard and concentration–response assessment of chemicals: the 1000 genomes high-throughput screening study. Environmental health perspectives, 123(5), 458 (2015). [PubMed: 25622337]
86. Low SK, Zembutsu H, Nakamura Y. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer science, 109(3), 497–506 (2018). [PubMed: 29215763]
87. Ingle JN, Schaid DJ, Goss PE et al. Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors. Journal of Clinical Oncology, 28(31), 4674 (2010). [PubMed: 20876420]
88. Chung S, Low S-K, Zembutsu H et al. A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. Breast Cancer Research, 15(5), R81 (2013). [PubMed: 24025145]
89. Godman B, Finlayson AE, Cheema PK et al. Personalizing health care: feasibility and future implications. BMC Medicine, 11(1), 179 (2013). [PubMed: 23941275] ** This is a very comprehensive review in precision medicine, including general considerations, pharmacogenomics, toxicity of drug therapies, biomarker tests, and new targeted therapies.

90. Keegan BMJTLN. Natalizumab for multiple sclerosis: a complicated treatment. Lancet Neurol, 10(8), 677–678 (2011). [PubMed: 21777819]
91. Kappos L, Bates D, Edan G et al. Natalizumab treatment for multiple sclerosis: updated recommendations for patient selection and monitoring. Lancet Neurol 10(8), 745–758 (2011). [PubMed: 21777829]
92. Mok TS, Wu Y-L, Thongprasert S et al. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. New England Journal of Medicine, 361(10), 947–957 (2009). [PubMed: 19692680]
93. Shah RR, Shah DRJBjocp. Personalized medicine: is it a pharmacogenetic mirage? British journal of clinical pharmacology, 74(4), 698–721 (2012). [PubMed: 22591598]
94. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. Journal of chemical information and modeling, 55(2), 263–274 (2015). [PubMed: 25635324]
95. Medina Marrero R, Marrero-Ponce Y, Barigye S et al. QuBiLs-MAS method in early drug discovery and rational drug identification of antifungal agents. SAR and QSAR in Environmental Research, 26(11), 943–958 (2015). [PubMed: 26567876]
96. Weidlich IE, Filippov IV, Brown J et al. Inhibitors for the hepatitis C virus RNA polymerase explored by SAR with advanced machine learning methods. Bioorganic & medicinal chemistry, 21(11), 3127–3137 (2013). [PubMed: 23608107]
97. Newby D, Freitas AA, Ghafourian T. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. European journal of medicinal chemistry, 90, 751–765 (2015). [PubMed: 25528330]
98. Jain N, Gupta S, Sapre N, Sapre NS. In silico de novo design of novel NNRTIs: a bio-molecular modelling approach. RSC Advances, 5(19), 14814–14827 (2015).
99. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, 43(6), 1947–1958 (2003). [PubMed: 14632445]
100. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research, 15(1), 3133–3181 (2014).
101. Singh H, Singh S, Singla D, Agarwal SM, Raghava GP. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. Biology direct, 10(1), 10 (2015). [PubMed: 25880749]
102. Mistry P, Neagu D, Trundle PR, Vessey JD. Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. Soft Computing, 20(8), 2967–2979 (2016).
103. Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms. Computers in biology and medicine, 56, 175–181 (2015). [PubMed: 25437231]
104. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527–1554 (2006). [PubMed: 16764513]
105. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. Molecular pharmaceutics, 13(5), 1445–1454 (2016). [PubMed: 27007977]
106. LeCun Y, Bengio Y, Hinton G. Deep learning. nature, 521(7553), 436 (2015).
107. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug discovery today, 22(11), 1680–1685 (2017). [PubMed: 28881183] ** This paper traces the evolution of machine learning development and discusses the insight into recently developed deep learning approaches and their applications in rational drug discovery.
108. Xue H, Li J, Xie H, Wang Y. Review of drug repositioning approaches and resources. International journal of biological sciences, 14(10), 1232 (2018). [PubMed: 30123072] ** This paper reviews three widely used drug repositioning approaches and summarizes 76 important resources about drug repositioning.

109. Sander J, Ester M, Kriegel H-P, Xu X. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169–194 (1998).
110. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications (ACM, 1998).
111. Wang W, Yang J, Muntz R. STING: A statistical information grid approach to spatial data mining. In: VLDB. (Ed.^(Eds) (1997) 186–195.
112. Ankerst M, Breunig MM, Kriegel H-P, Sander J. OPTICS: ordering points to identify the clustering structure. In: ACM Sigmod record. (Ed.^(Eds) (ACM, 1999) 49–60.
113. Chen B, Ding Y, Wild DJJPcb. Assessing drug target association using semantic linked data. PLoS computational biology, 8(7), e1002574 (2012). [PubMed: 22859915]
114. Zhu Q, Tao C, Shen F, Chute CG. Exploring the pharmacogenomics knowledge base (pharmgkb) for repositioning breast cancer drugs by leveraging web ontology language (OWL) and cheminformatics approaches In: Biocomputing 2014. (World Scientific, 2014) 172–182.
115. Consortium ICG. International network of cancer genome projects. Nature, 464(7291), 993 (2010). [PubMed: 20393554]
116. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine, 372(9), 793–795 (2015). [PubMed: 25635347]
117. Gaziano JM, Concato J, Brophy M et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. Journal of clinical epidemiology, 70, 214–223 (2016). [PubMed: 26441289]
118. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics, 19(4), 208 (2018).
119. Photopoulos J 13 countries to share 1 million genomes for research. (Ed.^(Eds) (bionews.orguk, 2018)
120. Robison RJ. The Genome War, Round Two. (Ed.^(Eds) (medium.com, 2015)
121. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AGJBib. Drug repositioning for orphan diseases. Briefings in bioinformatics, 12(4), 346–356 (2011). [PubMed: 21504985]
122. Wang Y, Kung L, Byrd TAJTF, Change S. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3–13 (2018).
123. Jensen MA, Ferretti V, Grossman RL, Staudt LMJB. The NCI Genomic Data Commons as an engine for precision medicine. 130(4), 453–459 (2017).
124. Bhise NS, Ribas J, Manoharan V et al. Organ-on-a-chip platforms for studying drug delivery systems. Journal of Controlled Release, 190, 82–93 (2014). [PubMed: 24818770]
125. Hotta A, Yamanaka SJArog. From genomics to gene therapy: induced pluripotent stem cells meet genome editing. 49, 47–70 (2015).
126. Dilsizian SE, Siegel ELJCcr. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current cardiology reports, 16(1), 441 (2014). [PubMed: 24338557]

Article highlights

- The accumulating drug knowledge bases, multi-omics data, and clinical data, comprise the cross-domain big data, facilitating the systematical drug discovery.
- The increasing informatics infrastructure now enables big data analysis to explore new drug therapeutics from multiple perspectives, such as genomics, proteomics, GWAS, pathway, EHR and pheWAS.
- Numerous well-defined and high-quality clinical phenotypic information available greatly facilitate the construction of *in silico* drug safety modeling in the early stage of drug discovery to screen low-toxicity compounds, to some extent, bypassing less-reliable rodent models.
- Rapidly developing machine learning techniques have drastically accelerated the big data-based drug discovery approach, further promoting the performance for both de novo drug discovery and drug repurposing.
- Incorporation of direct-to-consumer genetic testing and information technology industries has drastically strengthened the big data-based approach, further enhancing individualized precision medicine.

Table 1.

Resources for big data-driven therapeutic target identification.

Type	Focus	Resource	Assay/data type	Species	URL
Projects/initiatives	Genome, phenome, microbiome	Personal Genome Project (PGP)	DNA-seq, clinical phenotypes	Human	www.personalgenomes.org/harvard
	Genomic DNA variations across populations	1000 Genomes	DNA-seq	Human	www.1000genomes.org/data
	Genomic DNA variations across populations	HapMap	SNP array	Human	hapmap.ncbi.nlm.nih.gov/downloads/index.html.en
	Genomic DNA variations for rare disease and cancer	100,000 Genomes Project	DNA-seq, clinical information	Human	www.genomicsengland.co.uk
	Regulatory elements in genome	The Encyclopedia of DNA Elements (ENCODE)	ChIP/RNA/DNA-seq, DNA microarray	Various	www.encodeproject.org
	Multi-omic data in cancer	The Cancer Genome Atlas (TCGA)	DNA-seq, RNA-seq, methylation array, clinical phenotypes	Human	tcga-data.nci.nih.gov/tcga
	Multi-omic data in cancer	The International Cancer Genome Consortium (ICGC)	DNA-seq, RNA-seq, methylation array, clinical phenotypes	Human	dcc.icgc.org
	Somatic mutation mechanisms causing human cancer.	Cancer Genome Project at the Wellcome Trust Sanger Institute	DNA-seq, cell biology	Human	www.sanger.ac.uk/science/groups/cancer-genome-project
	Molecular signatures of gain/loss of genes	Library of Integrated Network-Based Cellular Signatures (LINCS)	Bead-array with informatic inference	Human	www.lincsproject.org/data/tools-and-databases
	Expression/DNA variants across organs	Genotype-Tissue Expression (GTEx)	RNA-seq, SNP array	Human	www.gtexportal.org/home
	Expression/DNA variants in cancer cell lines	Cancer Cell Line Encyclopedia (CCLE)	Expression array, targeted DNA-seq	Human	www.broadinstitute.org/ccle
	Presence/variations of microorganisms in human	Human Microbiome Project (HMP)	DNA-seq, reference genome, clinical meta-data	Microorganisms	hmpdacc.org/resources/tools_protocols.php
Data repository	Somatic DNA mutations in cancer	Catalogue of Somatic Mutations in Cancer (COSMIC)	Various omic types from literature/databases	Human	cancer.sanger.ac.uk/cosmic
	Various	Gene Expression Omnibus (GEO)	Various omic types with or without clinical phenotypes	Various	www.ncbi.nlm.nih.gov/geo
	Various	ArrayExpress	Various omic types with or without clinical phenotypes	Various	www.ebi.ac.uk/arrayexpress
	Expression/DNA variations, phenotypes	European Genome-phenome Archive	Various omic types, clinical phenotypes	Human	www.ebi.ac.uk/ega/home
	Annotated gene sets	Molecular Signature Database (MSigDB)	Gene expression, knowledgebase, genomic/genetic structural	Various	www.broadinstitute.org/msigdb
	Expression/DNA variations, phenotypes in cancer	Oncomine	Various omic types with or without clinical phenotypes	Human	www.oncomine.org/resource/login.html
	Multi-omic data in mouse models	Mouse Genome Informatics (MGI)	Various omic types from literature/databases	Mouse	www.informatics.jax.org
	Clinically relevant DNA variations	Clinical Genome Resource (ClinGen)	Various DNA variant assays, clinical phenotypes	Human	www.clinicalgenome.org
	Multi-omic data in mental illness	NIMH Repository and Genomics Resources (RGR)	biosamples, genetic, pedigree and clinical data	Human	www.nimhgenetics.org
	Cancer genomics data	cBioPortal for cancer genomics	DNA-seq, RNA-seq	Human	www.cbioportal.org

Table 2.

Resources for big data-driven drug identification.

Type	Focus	Resource	Assay/data type	Species	URL
Data repository	Chemical property/structure/biological function	ChemBank	Chemical properties, phenotypic readouts	Various	chembank.broadinstitute.org
		PubChem	Chemical properties, phenotypic readouts	Various	pubchem.ncbi.nlm.nih.gov
		Small Molecule Pathway Database (SMPDB)	Metabolomics pathway database	Human	smpdb.ca
		Human Metabolome Database (HMDB)	Human metabolites resouces	Human	www.hmdb.ca
		SCRIPDB	Chemical structure database	Various	dcv.uhnres.utoronto.ca/SCRIPDB/
		Binding Database (BindingDB)	Drug target database	Various	www.bindingdb.org/bind/index.jsp
		ChEMBL	Chemical properties, phenotypic readouts	Various	www.ebi.ac.uk/chembl
		Phamacogenomics Knowledgebase (PharmGKB)	Drug annotations, drug-gene association	Human	www.pharmgkb.org
		KEGG DRUG	Drug annotations, drug-gene association		www.genome.jp/kegg/drug
		DrugBank	Drug and drug target information	Various	www.drugbank.ca
		FDA Orange Book	List of FDA-approved drugs	Human	www.accessdata.fda.gov/scripts/cder/ob
	Protein properties	UniProt/Swiss-Prot	Protein sequence, structure, function, ontology	Various	www.uniprot.org
		Protein Data Bank (PDB)	Protein sequence, structure, function, ontology	Various	www.wwpdb.org
		Max-Planck Unifed Proteome Database (MAPU)	Protein sequence, structure, function, ontology	Human, mouse	mapuproteome.com
		ExPASy	Protein sequence, structure, function, ontology	Various	www.expasy.org
		Protein Information Resource (PIR)	Protein sequence, structure, function, ontology	Various	pir.georgetown.edu
		CATH	Classification of protein structures	Various	www.cathdb.info
		Human Protein Reference Database (HPRD)	Protein database	Human	www.hprd.org
		Plasma Proteome Database (PPD)	List of plasma/serum proteins	Human	www.plasmaproteomedatabase.org
	Protein-protein/chemical/genetic interaction	Database of Interacting proteins (DIP)	Experimentally determined protein-protein ineraction	Various	dip.mbi.ucla.edu/dip
		STITCH	Chemical-protein interaction	Various	stitch.embl.de
		BioGRID	Protein, genetic, chemical interactions/associations	Various	thebiogrid.org
		Therapeutic Target Database (TTD)	Protein, genetic, chemical interactions/associations	Human	database.idrb.cqu.edu.cn/TTD
		Potential Drug Target Database (PDTD)	Protein, genetic, chemical interactions/associations	Human	www.dddc.ac.cn/pdtd
Tools		NIH Small Molecule Repository	Compound libraries for screening	-	nihsmr.evotec.com/evotec
		ViCi	*in silico* ligand-based drug design	-	www.embl-hamburg.de/vici

Table 3.

Resources for computational prediction of drug toxicity.

Focus	Resource	Assay/Data type	URL
Computational prediction of drug toxicity	DEREK Nexus	Toxicological profile based on structure	www.lhasalimited.org/products/derek-nexus.htm
	ToxTree	Toxicological profile based on structure	toxtree.sourceforge.net
	HazardExpert	Toxicity profile based on toxic fragments	www.compudrug.com/hazardexpertpro
	TOPKAT	Toxicological profile based on structure. Predicted ADMET properties.	accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html
	CASE Ultra	Toxicological profile based on structure	www.multicase.com/case-ultra
	OECD QSAR	Toxicological profile based on structure	www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm
Database of bioactive molecules and targets	ChEMBL	Chemical properties, phenotypic readouts	www.ebi.ac.uk/chembl
	DrugBank	Drug and drug target information	www.drugbank.ca
	Drugs@FDA Database	FDA-approved drugs	www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
	PubChem	Chemical properties, phenotypic readouts	pubchem.ncbi.nlm.nih.gov
	SWEETLEAD	Chemical structure database	simtk.org/home/sweetlead
	The NCGC Pharmaceutical Collection (NPC)	Chemical structure database	tripod.nih.gov/npc
	Chemical Entities of Biological Interest (ChEBI)	Small compound database	www.ebi.ac.uk/chebi
Toxicology databases	SIDER	Adverse drug reaction database	sideeffects.embl.de
	Comparative Toxicogenomics Database	Chemical-protein and chemical-phenotype interactions	ctdbase.org
	Environmental Protection Agency Aggregated Computational Toxicology Resource (ACToR)	Chemical toxicity database	actor.epa.gov
	FDA Adverse Event	Adverse drug reaction database	www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects
	OpenTox	Multiple toxicological resources (data, computer models, validation and reporting)	www.opentox.org
	Phamacogenomics Knowledgebase (PharmGKB)	Drug annotations, drug-gene association	www.pharmgkb.org
	T3DB	Toxin-target database	www.t3db.ca
	TOXNET	Toxicology resources	toxnet.nlm.nih.gov
	ToxBank	Toxicology resources	www.toxbank.net
	SuperToxic	Toxicology resources	bioinformatics.charite.de/supertoxic/
	Aggregated Computational Toxicology Resource (ACToR)	Chemical toxicity database	actor.epa.gov/actor/home.xhtml
	eTOX	Toxicological data and models to support toxicity prediction	www.etoxproject.eu

Focus	Resource	Assay/Data type	URL
	Hazardous Substances Data Bank (HSDB)	Toxicology of potentially hazardous chemicals	toxnet.nlm.nih.gov/newtoxnet/hsdb.htm
	Haz-Map	Occupational toxicology database	hazmap.nlm.nih.gov
	Chemical Effects in Biological Systems (CEBS)	Chemical effects database	tools.niehs.nih.gov/cebs3/ui/
Chemical property/structure	Chembank	Chemical properties, phenotypic readouts.	chembank.broadinstitute.org

Table 4.

Resources for commercial cloud services.

Service	Platform	URL
Software as a service (SaaS)	Amazon Web Services	www.aws.amazon.com
	Google Cloud Platform	cloud.google.com
	Microsoft Azure	azure.microsoft.com
	IBM Cloud	www.ibm.com/cloud/
	Alibaba Cloud	www.alibabacloud.com
Infrastructure as a service	DNAnexus	www.dnanexus.com
	Illumina BaseSpace Sequence Hub	basespace.illumina.com
	Seven Bridges	www.sevenbridges.com/platform
	Globus Genomics	globusgenomics.org