EXPERT REVIEWS

Big data: the next frontier for innovation in therapeutics and healthcare

Expert Rev. Clin. Pharmacol. 7(3), 293–298 (2014)

Naiem T Issa[1], Stephen W Byers[1,2] and Sivanesan Dakshanamurthy*[1,2]

[1]*Department of Oncology, Lombardi Cancer Center, Georgetown University Medical Center, Washington, DC USA*
[2]*Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington, DC USA*
**Author for correspondence: Tel.: +1 202 687 2347 sd233@georgetown.edu*

Advancements in genomics and personalized medicine not only effect healthcare delivery from patient and provider standpoints, but also reshape biomedical discovery. We are in the era of the '-omics', wherein an individual's genome, transcriptome, proteome and metabolome can be scrutinized to the finest resolution to paint a personalized biochemical fingerprint that enables tailored treatments, prognoses, risk factors, etc. Digitization of this information parlays into 'big data' informatics-driven evidence-based medical practice. While individualized patient management is a key beneficiary of next-generation medical informatics, this data also harbors a wealth of novel therapeutic discoveries waiting to be uncovered. 'Big data' informatics allows for networks-driven systems pharmacodynamics whereby drug information can be coupled to cellular- and organ-level physiology for determining whole-body outcomes. Patient '-omics' data can be integrated for ontology-based data-mining for the discovery of new biological associations and drug targets. Here we highlight the potential of 'big data' informatics for clinical pharmacology.

Keywords: big data • clinical pharmacology • personalized medicine • systems medicine • therapeutics

The digital revolution in healthcare is now. The amount of data is exploding from basic science to clinically based genomics and personalized medicine, and continues to evolve in healthcare at both the population and the individual levels. Clinical phenotypes are being described more quantitatively and biochemically using genomics, transcriptomics, proteomics and metabolomics [1–3]. Collecting and analyzing such large data sets, coined 'big data', will become key to new healthcare innovations. In anticipation, the USA and European Commission announced a national 'Big Data Initiative' for policy preparedness [4].

Advancements in big data analytics not only affect healthcare delivery from patient and provider standpoints but also hold promise to reshape biomedical discovery. For example, decoding a single human genome originally took a decade to process, but with the advent of 'Big Science', modern DNA sequencing and informatics approaches can achieve this within a week. With respect to clinical pharmacology, a big data initiative by Medco recently helped uncover that the simultaneous use of clopidogrel (Plavix) and proton pump inhibitors is associated with increased risk of adverse cardiovascular events [5]. 'Systems medicine' has brought with it a slew of technologies and novel perspectives to analyze disease as an interconnected synergistic system of working parts. Clinical pharmacology is transitioning to reflect this interconnectedness, where the traditional unidirectional movement of 'bench-to-bedside' is now a bidirectional process that depends on both end points to achieve novel therapeutics and better understanding of effects caused by current pharmaceuticals. The following sections provide insight into the available resources and techniques being applied in this transformation.

Big data in clinical discovery

New-age drug discovery approaches encompass a wide range of big data analytics, from high-throughput cellular and protein-binding assays to chemoinformatics-driven databases. Many of these databases are undergoing extensive improvements to become central hubs for the integration of biological and physicochemical information (Table 1). In addition, new patterns and associations are being discovered using

 10.1586/17512433.2014.905201 ISSN 1751-2433

Table 1. Table of open-access resources available to the public for bioinformatics and chemoinformatics in drug discovery and toxicology.

Category	Database	Description
Drug	BindingDB	Binding affinities of small, drug-like molecules to protein targets
	ChEBI	Small chemical compounds containing structural, nomenclature and ontology information
	ChemBank	Biomedical measurements derived from cell lines treated with small molecules
	ChEMBL	Manually curated bioactive molecules with drug-like properties maintained by the EBI
	DrugBank	FDA-approved and experimental drugs with drug target, bio- and chemoinformatic data
	PharmGkb	Pharmacogenomic-focused genetic, molecular, cellular and clinical data for drugs
	SuperTarget	~7300 drug–target associations with ~5000 manually annotated
	Therapeutic Target Database	Known therapeutic protein targets with pathway information and corresponding drugs
	ZINC	~21 million compounds that are commercially available and prepared for virtual screening
Disease	National Organization of Rare Diseases	NORD contains information on rare human diseases
	Online Mendelian Inheritance of Man	Catalog of human genes and genetic disorders maintained by the Johns Hopkins University
Protein–protein/–gene/ –other interactions	BioGRID	~730,000 raw protein and genetic interactions from major model organisms
	DGIdb	Drug gene interaction database curated from multiple well-established databases
	ExPASy STRING	Known and predicted protein–protein interactions from experimental repositories and computational methods
	MatrixDB	Interactions between extracellular proteins (i.e., collagen and laminins) and polysaccharides
	MINT	Molecular interaction database focusing on experimentally validated protein–protein interactions
	Database of Interacting Proteins	Manual and computational curation of experimentally determined protein–protein interactions
Genomics	Gene Expression Omnibus	Public functional array- and sequence-based genomics data repository
	Oncomine	Cancer microarray database that can be subdivided by treatment, patient survival and other demographics
	The Cancer Genome Atlas	Large-scale genome sequencing platform for multiple cancers led by the NCI and the NHGRI
	UCSC Cancer Genome Browser	Interactive annotated cancer genome-browser website hosted by the University of California, Santa Cruz
	G-DOC	Broad collection of bioinformatics and systems biology tools for analysis and visualization of four major 'omics' types: DNA, mRNA, microRNA and metabolites
Proteomics	dbDEPC	Database of differentially expressed proteins in human cancer
	GeMDBJ Proteomics	Clinical and cell line protein LC-MS/MS and 2D-difference gel electrophoresis for expression levels
	Plasma Protein Database	Initiative of the Human Proteome Organization to characterize human plasma and serum proteome
	PRIDE	Centralized standards-compliant mass spectrometry proteomics and post-translational modifications
	The Human Protein Atlas	Immunohistochemistry-based protein expression profiles of various human tissues, cancers and cell lines
	UniProt	Comprehensive protein sequence and annotation data
Metabolomics	BiGG	Genomic-based reconstruction of human metabolism for systems biology simulation and flux modeling
	HMDB	Human small molecule metabolites with associated chemical, clinical and molecular biology information
	HumanCyc	Human metabolic pathway/genome bioinformatics database constituting over 28,000 genes
	SMPDB	Small molecule pathway database with >400 unique human pathways not found in other databases

Table 1. Table of open-access resources available to the public for bioinformatics and chemoinformatics in drug discovery and toxicology (cont.).

Category	Database	Description
Toxicology	Chemical Effects in Biological Systems	Developed by the National Institute of Environmental Health Sciences to house toxicology studies
	Comparative Toxicogenomics Database	Curation of chemical–gene, chemical–disease, and gene–disease associations and constructs networks
	EPA ACToR	Aggregated Computational Toxicology Resource to query multiple EPA chemical toxicity databases
	FDA Adverse Event Reporting System	Adverse event and medication error reports submitted to the FDA for post-market safety surveillance
	OpenTox	Interoperable framework for predictive toxicology and community platform for creation of applications
	SIDER	Adverse drug reactions on marketed medicines extracted from public documents and package inserts
	T3DB	~37,000 pollutant–, pesticide– and food toxin–target associations with 50 related bioinformatics data fields
	TOXNET	Integrated database system of hazardous chemicals, toxic releases and environmental health by the NLM
Structural biology/ ontology/enzymology/ pathways	BRENDA	Comprehensive repository for molecular and biochemical information of enzymes classified by the IUBMB
	Gene Ontology	Gene and gene product annotations across multiple species with supported tools for analytics
	Ingenuity	Web-based applications for analyzing genomic and pathway data
	KEGG	Five databases that connect molecular interaction networks to recapitulate the human biological system
	RCSB PDB	The Protein Data Bank containing protein crystal structures and ligand binding affinities
	Reactome	Manually curated and peer-reviewed pathway database with cross-references and visualization tools
	SCOP	Structural classifications of proteins deposited in the PDB using a tree-like hierarchy

external computational tools that mine the data. Thus, new therapeutic targets, drug–target associations and drug repurposing hypotheses can be established by reassessing this large amount of data through more integrative approaches. For example, our group previously devised computational platform entitled 'Train, Match, Fit, Streamline' (TMFS) for predicting empirical drug-target signatures and establishing repurposing hypotheses [6]. While it is important to establish these signatures from the low-level biochemical standpoint, their full clinical potential remains unfulfilled until placed into a network perspective that places those predictions within the context of systems medicine using the aforementioned databases. We are currently interfacing TMFS with these public databases for big-data network pharmacological applications [Issa et al., Unpublished data].

Additional interesting general outcomes from these approaches include the observations that the chemical space spanned by drugs does not necessarily correlate with biological activity [7], and that singular targets may not be sufficient to alter diseases [8]. It naturally follows that combining network systems biology with pharmacology could unveil critical associations or even combinatorial therapeutic strategies for recalcitrant diseases. From the genomics standpoint, for example, Lamb *et al.* designed the Connectivity Map, a repository for cell type-specific drug-induced gene expression changes [9]. If diseases are associated with unique gene signatures, the current thinking is that a drug that induces an inversely correlated signature should be therapeutic. By looking broadly into expression patterns, the bias of choosing canonical or literature-rich pathways for therapeutic targeting is bypassed. An interesting finding from the Connectivity Map is the ability of rapamycin, an mTOR inhibitor, to overcome dexamethasone resistance in the CEM-c1 lymphoid cell line. There is an active clinical trial assessing the repurposing potential of rapamycin in children exhibiting dexamethasone-resistant acute lymphoblastic leukemia [10].

Big data in systems medicine & pharmacology

Genomics has been particularly useful in oncology. Genetically, tumors are intrinsically heterogeneous and undergo further genetic changes upon exposure to stresses (i.e., hypoxia) or phenotypic changes (i.e., epithelial-to-mesenchymal transformation). These changes often allow tumors to escape current pharmacological interventions. To better understand these changes, and perhaps their predictability, notable efforts such as The Cancer Genome Atlas (TCGA), Oncomine and the University of California Santa Cruz Cancer Genome Browser document genome-wide changes of various cancers (Table 1). Investigators are able to access this information to analyze differential expression patterns

across cell lines, disease progression states, treatment, etc. With big data, subtleties have been uncovered to provide insight into resistance mechanisms and new pharmacological targets. A recent example is characterizing acquired resistance to vemurafenib in melanoma, in which MAPK pathway reactivation and PI3K-PTEN-AKT activation were discovered to be two core resistance mechanisms [11]. This information aids in devising second-stage or combinatorial strategies in anticipation of resistance in BRAF V600E-mutant melanoma patients.

While genomics has its strengths, protein expression patterns and metabolite composition provide more nuanced insight into end point phenotypes. As techniques for quantifying and characterizing proteomes and metabolomes become cheaper and more sensitive, there are more initiatives employing these methods for disease characterization and therapy. For cancer, the National Cancer Institute released in September 2013 the first public proteomic data of colorectal tumor samples previously analyzed by the TCGA – the first complementation of proteomic and genomic data on the same tumors. This coupling is instrumental in establishing genotype–phenotype associations and elucidating precisely which signaling pathways contribute to pathogenesis and serve as avenues for therapy. Similarly, for cancer, metabolites not only serve as diagnostic aids but also as potential biomarkers that demarcate disease progression, response to therapies and new therapeutic targets [12,13]. As biological, chemical and clinical data continue to increase at an exponential pace, computer-assisted data mining platforms that integrate the information and convey it in meaningful contexts tailored to appropriate investigators provide an invaluable resource for drug discovery and repurposing.

Big data & toxicity prediction

Unintended deleterious effects continue to hamper the success of drug development, especially during late-phase clinical trials and post-market surveillance. These effects compromise health outcomes and incur losses in research and development. Toxicity is attributed to a variety of factors, but common culprits include dosing, patient-specific sensitivities (i.e., allergic responses, cytochrome pharmacogenomics) and prolonged clinical symptom detection time. Multiple databases exist that couple drugs with side effects and serve as the basis for informatics and network studies that aim to predict side effects for new drugs (Table 1). These efforts attempt to associate chemistry/structure or biological pathways with molecular and clinical side effects.

While current databases are immensely useful for studying observed drug–side effect associations, still many rare occurrences or non-intuitive phenotypic links are not captured. Data mining efforts are attempting to uncover these missed associations by semantically linking drugs to clinical or cellular side effects and then discovering the causative genes and pathways. For example, Xu and Wang developed a rank-prioritized cancer-specific drug-side effect lexicon from the entire MEDLINE corpus and found that cancer drugs that share side effects tend to have overlapping gene targets and indications [14]. More recently, van Haagen *et al.* utilized 'concept recognition' to mine MEDLINE for biomedical concepts and found that generic concepts, such as 'diagnosis' or 'etiology', are crucial for inferring plausible protein–protein interactions [15]. These examples are just a sample of many that are adopting whole-literature semantic data mining techniques for building veritable associations that are useful not just for predictions or assessments in toxicological systems pharmacology, but also for therapeutic drug discovery and repurposing.

While reanalyzing current data is critical, it is equally important to clarify the temporality of toxicity and tailor current technologies to characterize the toxicity time course. Toxicity is presently observed as a clinical phenomenon with well-defined symptoms. However, it is vital to realize that these symptoms develop in relation to time, dose and mechanism of exposure, while sub-clinical toxicity at the cellular level occurs long before symptom detection. Thus genomic, proteomic and metabolomic profiles of different organs and body fluid compartments are likely to change over the course of toxicity and indicate different stages of progression. Assessing these profiles could prove beneficial for patient-specific monitoring of drug-induced toxicity, pre-treatment prognosis of toxicity and post-toxicity treatment options. Current 'omics' applications are characteristically epitomized by Gao [16] and Clayton [17]. The work of Gao *et al.* eschews the coming of toxicoproteomics, where systematic protein quantitation enables pathway elucidation, protein–protein interaction and protein subcellular localization as exposure outcomes for drugs, toxicants and other environmental stressors [16]. Similarly, Clayton *et al.* developed personalized pharmacometabolomics, where pre-dose urinary metabolite profiles are predictive of drug-induced hepatocellular toxicity [17]. These cases are not reliant on the actual clinical symptomology: they can characterize toxic effects to high resolution in any tissue or body compartment at any time during any intervention. Toxicity, like disease, begins as a molecular phenomenon before clinical detection, and the combination of big data, high-resolution, molecular-level informatics with clinical outcomes greatly benefits pharmacological developments.

Big data & electronic medical records

An underutilized source of individual patient phenotypes is electronic medical records (EMRs). EMRs are rich with clinical data that chart patient progression with respect to disease and medications with other demographic information. Family history, diet, medications and occupational exposures are just some of the documented information and could be invaluable in discerning uniquely observed treatment or toxicity effects. Furthermore, the combinatorial aspect of assessing outcomes in polypharmacology becomes unrealistic in a controlled clinical trial setting but can be viably studied using EMRs.

Some investigator groups have peered into EMR potentiality for drug development and safety. For instance, Hanauer *et al.* extracted data from clinical problem summary lists for 327,000 patients and discovered novel disease association pairs using the Molecular Concept Map tool [18].

One novel relationship they reported is between osteoarthritis and granuloma annulare, both of which are treated by niacin. Associated diseases may have common pathogenic pathways with shared drug targets, thus aiding in target discovery as well as drug repositioning. With respect to pharmacovigilance, Castro *et al.* utilized EMRs to confirm that the antidepressant citalopram is associated with QT interval prolongation [19].

EMR information is uniquely positioned to aid in the discovery of new therapeutic targets when coupled with patient-derived 'omics' data. In other words, combining genotype–phenotype biomedical information with environmental and social contributors would provide a holistic systems view of a patient and highlight patient-specific changes for personalized pharmacology. To date, no such platform exists as both EMR standardization and 'omics' translation to clinical medicine are yet to be universally adopted. We anticipate that this merger will occur as clinical breakthroughs continue to emerge from each space.

Expert commentary & five-year view

It is clear that 'multi-omics' big data are mainstay in current research efforts and will soon be the case in the clinical setting. Utilizing such technologies and strategies would help improve the efficiency of research and clinical trials, identify and develop new effective medicines more quickly and build new tools for physicians to meet the promise of personalized medicine. The coming ubiquity of these technologies coupled with the need to approach healthcare from an interconnected systems standpoint will lead to a new era of health data acquisition with new difficulties to consider, such as data structure and management, data privacy and data analytics. Corporate efforts and start-ups are currently engaging the wider community encompassing patients, scientists and physicians to make these technologies more efficient and affordable for individualized use. One example is 23andMe, which offers to sequence the DNA of a consumer for $99 and provide raw genetic data. Proteomics is also becoming more industrialized, as evident by new companies such as Applied Proteomics, Inc.

With the role of the 'e-Patient' being not only a consumer of healthcare but also a driver of discovery, patient data access is currently a hotly debated topic [20]. Additionally, there will be a need to seamlessly interface data production with patient-specific health records. We envision the development of specialized software that directly interacts with EMRs that not only deposit patient-generated data but also simultaneously encrypt it in compliance with health privacy laws. This process could lead to a tiered system, modeled after the TCGA, for example, where data access to a particular tier is limited to a specific audience, thereby promoting involvement across both patients and clinician researchers. Moreover, data centralization will be a key initiative for current and future taskforces to avoid the fragmentation of the extremely large set of data and promote both national and international collaborative efforts. This centralization would allow for integrative data analytics using high-performance computing establishments and cloud platforms. Taking heed of these needs early is essential for establishing streamlined data acquisition, retrieval, analysis, dissemination and ultimately discovery.

Combining these technologies will advance modern clinical pharmacology. Patient-specific pharmacodynamics and pharmacokinetics can be elucidated, allowing for individualized therapeutic indices, combinatorial approaches and novel avenues for circumventing drug resistance. First-line therapies and their adjustments will be tailored to time- and therapy-dependent changes in the patient epigenome, transcriptome, proteome and metabolome. Furthermore, emerging technologies such as implantable 'lab-on-a-chip' diagnostic devices will enable real-time monitoring of patient physiology and even provide immediate point-of-care clinical decision support based on the aforementioned patient-specific data and mining of the available open-access databases via wi-fi, Bluetooth or cellular networks. Over the course of the next 5 to 10 years, current clinical protocols will become antiquated as the field progresses toward big-data-driven, evidence-based standards.

Conclusion

In summary, biological, chemical and clinical repositories are becoming core initiatives of multiple academic, governmental and industry big-data and 'omics' ventures. In order for big-data approaches to revolutionize healthcare, centralization and standardization of EHR collection is essential to facilitate collaboration and effective mining of the data to enable plausible pharmacological discoveries. High-throughput automated data acquisition methods are being implemented in every sector of the biosciences, and the rate at which the data are generated outpaces their analysis. The limiting factor in the application of these data is connecting patterns and idiosyncrasies to disease processes and therapeutic outcomes with biological plausibility. It is possible that supercomputers and cloud-based parallel high-performance computing (i.e., IBM's Watson and Oracle's cloud platform, respectively) could derive veritable biological hypotheses that are continuously reformulated for accuracy as the data are being constantly fed. As big data in the biosciences continues to grow, clinicians, scientists and IT specialists must approach the information in an interdisciplinary fashion to promote standardization and a new epoch of discoveries in the realm of clinical pharmacology.

Financial & competing interests disclosure

Authors are supported in part by the National Institutes of Health grant CA170653 and Lombardi Cancer Center CCSG grant NIH-P30 CA51008. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Key issues

- Big data in medicine is quickly transitioning from the research sector to the public community space, thus driving new biomedical discoveries, reassessment of current healthcare policies and reshaping of clinical practice.
- Creating effective individualized healthcare programs can be achieved with the implementation of big data in systems and clinical medicine.
- We provide insight into the influence of big data, in the form of '-omics' and electronic medical records, on clinical pharmacology, as well as the evolving computational tools and platforms used for analytics.
- The current perspective envisions not only the discovery of novel drug–gene signatures and gene networks through big data but also the refined characterization of pharmaceutical-induced toxicity.
- In conclusion, big data further enables systems and clinical pharmacological approaches at an integrative and holistic level that can be applied to an individual for patient-specific treatment and healthcare maintenance.

References

1. Robinette SL, Holmes E, Nicholson JK, Dumas ME. Genetic determinants of metabolism in health and disease: from biochemical genetics to genome-wide associations. Genome Med 2012;4:30
2. Heidecker B, Hare JM. The use of transcriptomic biomarkers for personalized medicine. Heart Fail Rev 2007;12:1-11
3. Liotta LA, Kohn EC, Petricoin EF. Clinical proteomics: personalized molecular medicine. JAMA 2001;286:2211-14
4. Obama administration unveils "Big Data" initiative: announces $200 million in new R&D investments. The White House. Big Data Public Private Forum; Washington, DC: 2012. Available from: www. cordis. europa.eu [Last accessed 05 March 2013]
5. Kreutz RP, Stanek EJ, Aubert R, et al. Impact of proton pump inhibitors on the effectiveness of clopidogrel after coronary stent placement: the clopidogrel Medco outcomes study. Pharmacotherapy 2010;30: 787-96
6. Dakshanamurthy S, Issa NT, Assefina S, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012;55:6832-60
7. Root DE, Kelley BP, Stockwell BR. Global analysis of large-scale chemical and biological experiments. Curr Opin Drug Discov Devel 2002;5(3):355-60
8. Barabasi A, Oltvai Z. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101-13
9. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929-35
10. ClinicalTrials.gov identifier: NCT00874562. Rapamycin in relapsed acute lymphoblastic leukemia. Available from: http:// clinicaltrials.gov/show/NCT00874562
11. Shi H, Hugo W, Kong X, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov 2014;4:80-93
12. Ganti S, Weiss R. Urine metabolomics for kidney cancer detection and biomarker discovery. Urol Oncol 2011;29:551-7
13. Aboud O, Weiss R. New opportunities from the cancer metabolome. Clin Chem 2013;59:138-46
14. Xu R, Wang Q. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug–side effect relationships from the literature. J Am Med Inform Assoc 2014;21:90-6
15. van Haagen HH, 't Hoen PA, Mons B, Schultes EA. Generic information can retrieve known biological associations: implications for biomedical knowledge discovery. PLoS One 2013;8:e78665
16. Gao Y, Holland R, Yu L. Quantitative proteomics for drug toxicity. Brief Funct Genomic Proteomic 2009;8:158-66
17. Clayton TA, Lindon JC, Cloarec O, et al. Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature 2006;440:1073-7
18. Hanauer D, Rhodes D, Chinnaiyan A. Exploring clinical associations using '-omics' based enrichment analyses. PLoS One 2009;4:e5203
19. Castro VM, Clements CC, Murphy SN, et al. QT interval and antidepressant use: a cross sectional study of electronic health records. BMJ 2013;346:f288
20. Lunshof J, Church G, Prainsack B. Raw personal data: providing access. Science 2014;343:373-4