HHS Public Access

Author manuscript

Clin Pharmacol Ther. Author manuscript; available in PMC 2017 March 01.

Published in final edited form as:

Clin Pharmacol Ther. 2016 March ; 99(3): 250–254. doi:10.1002/cpt.322.

Big Data Transforms Discovery-Utilization Therapeutics Continuum

SA Waldman[1] and **A Terzic**[2]

[1]Department of Pharmacology and Experimental Therapeutics, Division of Clinical Pharmacology, Department of Medicine, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

[2]Mayo Clinic Center for Regenerative Medicine, Divisions of Cardiovascular Diseases and Clinical Pharmacology, Departments of Medicine, Molecular Pharmacology and Experimental Therapeutics and Medical Genetics, Mayo Clinic, Rochester, Minnesota, USA

Abstract

Enabling omic technologies adopt a holistic view to produce unprecedented insights into the molecular underpinnings of health and disease, in part, by generating massive high-dimensional biological data. Leveraging these systems-level insights as an engine driving the healthcare evolution is maximized through integration with medical, demographic, and environmental datasets from individuals to populations. Big data analytics has accordingly emerged to add value to the technical aspects of storage, transfer, and analysis required for merging vast arrays of omic-, clinical- and eco-datasets. In turn, this new field at the interface of biology, medicine, and information science is systematically transforming modern therapeutics across discovery, development, regulation, and utilization.

> "…a man's discourse was like to a rich Persian carpet, the beautiful figures and patterns of which can be shown only by spreading and extending it out; when it is contracted and folded up, they are obscured and lost"
>
> Themistocles quoted by Plutarch AD 46 – AD 120

Like the tapestry in Plutarch's quote, we can only comprehend the intricate patterns that constitute wellness and disease by spreading out and extending the multi-dimensional components that form the fabric of these processes. Implied in this self-evident concept is the ability to collect the relevant data, deconvolute that data into comprehensible elements, reassemble these elements into distinguishable patterns, and provide this new knowledge in a form that is readily accessible to end-users, including patients, practitioners and regulators.[1] The emergence of enabling medical technologies has revolutionized our ability to precisely define the detailed characteristics of individuals in sickness and health. Omic technologies offer a view of organization and function at the level of integrated molecular systems while next generation imaging imparts structure to those systems at cell, tissue,

Correspondence: Scott A. Waldman, MD, PhD, Thomas Jefferson University, 132 South 10th Street, 1170 Main, Philadelphia, PA 19107, scott.waldman@jefferson.edu and Andre Terzic, MD, PhD, Mayo Clinic, 200, First Street SW, Stabile 5, Rochester, MN 55905, terzic.andre@mayo.edu.

Financial Disclosures: The authors have no relevant disclosures.

organ and organismal levels.[2] Beyond these biological determinants, environmental elements that provide the context for molecular structure and function and, ultimately, shape pathobiology are memorialized in the longitudinal electronic health record (EHR).[3] Together, these biological and environmental data elements encode the information that predicts wellness, identifies disease risk, personalizes healthcare interventions, and prevents untoward adverse therapeutic events.

While these individual data elements form the matrix that defines the mechanisms underlying health and disease, a complete picture of these processes emerges only from their integration. Like a painting created in the style of the 19th century Pointillism technique, the entire picture only comes into view when one steps away from the canvas and coalesces the individual dots into a coherent image. In the context of biological and clinical data, the full picture of (patho)physiology emerges when these elements are integrated across individual patients and populations. The attendant challenges associated with this necessary data integration can be appreciated by considering the sheer magnitude of the task. In 2012, the worldwide digital healthcare data burden was estimated to be ~500 petabytes and is expected to reach 25,000 petabytes (~10^{19} kB) in 2020.[4] For comparison, the human brain stores ~2 petabytes of data while the largest single data storage facility is ~100 petabytes. In that context, it has become easier and cheaper to generate data than to store, integrate and analyze it.[5]

This avalanche of high dimensional data at the interface of biology, medicine, and healthcare delivery holds the potential to transform the therapeutics continuum of discovery, development, regulation, and utilization (DDRU).[6] As highlighted in the Commentary by Schneweiss, this informational nexus is poised to provide unprecedented insights into the pathobiology of disease, transform drug discovery and development, and revolutionize the ability of regulatory agencies to maintain the highest standards of drug safety, all focused on providing the best care precisely tailored to each individual patient.[7] However, these large and complex data sets are difficult to process using common database management tools or traditional data processing applications, especially with respect to data capture, storage, searching, sharing, integration and analysis. While the goal is to extract insights from complex, noisy, and heterogeneous data sets, barriers have included the speed of data handling, curation and the veracity of the data, the sheer volume of data, and the heterogeneity of data to be integrated.[7, 8] To address these challenges, big data analytics has emerged as a new discipline innovating the tools, processes and procedures that create, manipulate, manage and integrate very large heterogeneous data sets, to generate value from the whole that could not be appreciated form the sum of the individual parts.[7]

The potential for big data to transform paradigms of disease pathobiology is exemplified by the electronic health record (EHR), which in aggregate across the population represents an extremely large collection of information generated in routine clinical care.[4-6, 9] These datasets are challenging to use because they are heterogeneous, representing digital data as well as unstructured information, for example clinical notes. To optimize their utility, a new generation of technologies and architectures has emerged to extract value from large volumes of complex heterogeneous datasets through high-velocity capture, discovery and analysis.[5] Analytic tools to cull this information from these large collections of unstructured

data include artificial intelligence, natural-language processing, pattern recognition and machine learning.[5] In that context, in their review, Roden and Denny describe how coupling EHRs to genomic datasets specifically enable discovery of genotype-phenotype associations which, in turn, can then be implemented through EHRs to individualize patient care.[9] They highlight the global character of this effort, which includes their Electronic Medical Records and Genomics (eMERGE) Network, as well as the Veterans Administration's Million Veterans Program, the Kaiser-Permanente GERA program, the UK Biobank, and the Icelandic deCODE resource.[9, 10] Beyond their value in discovering common genetic loci associated with human disease through genome-wide association studies (GWAS), these resources also can be exploited to identify rare genetic variants with large effect sizes, pleiotropic effects of common and rare genetic variants, and potential drug targets.[11] One obstacle to the utility of EHRs for discovery research has been the ability of these databases to accurately identify clinical phenotypes that could be used to assemble true case and control cohorts to support meaningful genotype-phenotype correlations.[12] Indeed, for common diseases, where datasets could include hundreds of thousands of subjects, electronic algorithms have been developed to overcome this obstacle and extract true cases and control subjects, including the eMERGE's Phenotype Knowledgebase (PheKB.org) and i2b2 (informatics for integrating biology and the bedside).[9-11] Employing these approaches, drug response and adverse drug reaction phenotypes can be readily identified. Importantly, beyond these genotype-phenotype associations, which typically start with a defined disease (phenotype) to explore genomic associations, the constellation of phenotypes represented within the collective EHR – the EHR phenome – can be interrogated for genomic associations in phenome-wide association studies (Phe-WAS).[9]

The foregoing discussion underscores the potential for big data analytics as a resource for discovery of new molecular associations, disease pathways, and pathophysiological mechanisms. This is especially true in the context of integrating medical databases like the EHR and clinically-annotated omic databases that associate disease phenotypes with molecular features like genomics, epigenomics, transcriptomics, proteomics and metabolomics. As highlighted in the review by Chen and Butte, such databases have been constructed and are publically available to support an emerging in silico approach to drug discovery and development.[13] For example, transcriptomic analysis using data mining revealed that expression of the protein MTBP was significantly elevated in breast cancer samples compared to normal breast tissues and associated with poor survival.[13] Indeed, this gene product could be used to stratify breast cancer patients into clinically relevant subgroups and might represent a new therapeutic target in these populations.[13] Similarly, analysis of databases containing genomic characteristics of thousands of tumors revealed >400 new defects that serve as driver mutations that were previously unrecognized.[13] Further, mapping these driver genes to drug databases, including ChEMBL and ClinicalTrials.gov, revealed that >70% of patients could benefit from novel agents in clinical development.[13] Beyond discovery, these in silico approaches also can be used to assess target druggability, through an integrative analysis of protein function, homology to targets of approved drugs, three-dimensional structure, and the existence of published active small molecules.[13] Moreover, these analytic approaches can be employed to compare similarities across different diseases, and the different drugs used to treat them, to develop new

indications for existing agents through the emerging approach of computation drug repositioning.[13] Together, these considerations highlight the potential of big data analytic approaches to transform the science of drug discovery and development.[6, 13]

Beyond discovery and development, big data analytics is revolutionizing the safety of therapeutics at the level of regulation and utilization. As outlined in the Commentary by Harpaz, DuMochel and Shah, pharmacovigilance currently depends on spontaneous adverse event reporting from drug manufactures, health care professionals, and patients.[14] While this type of reporting is essential to post-marketing surveillance, and effective at detecting ADRs, it is a passive system fraught with delays in detecting and reporting, and a substantial number of ADRs remain unreported.[14] Big data analytics offers an unprecedented solution to improving pharmacovigilance, providing unique mechanisms for adverse event detection and evaluation. Some of the data sources to support this emerging field have been described earlier, for example the very large collections of information in the EHR.[7, 9] In that context, the EHR is the backbone for the FDA's Sentinel Initiative, described in the Commentary by Ball, Rob, Anderson, and Dal Pan, which is creating a national network of databases to prospectively monitor the safety of drugs and rapidly respond to emerging risks.[14-16] This Sentinel System currently comprises 18 partners, contains data on >170 million patients, and is earmarked to expand.[14, 16] Surprisingly, another evolving source of vast amounts of relevant information is social media.[5, 14] This real-time source of information includes health forums, social networks, and online patient communities, with posts typically occurring proximal to the time an event occurs.[5, 14, 15, 17, 18] One example is the algorithm used by Google to track diseases, Google Trends, which uses geospatial mapping to sift through enormous amounts of real-time data vast quantities of information to identify clinically-relevant population-level events.[5] For example, Google Trends can identify peaks in search requests for terms like ‘flu symptoms’ and ‘flu treatments’ to identify an imminent disease outbreak in a geographic region even before patients begin to task the regional health system.[5] This example highlights the opportunities offered by social media for adverse event surveillance that is global and real-time, to achieve the earliest detection.

While big data analytics will transform every facet of the DDRU continuum, it is not without significant challenges. There are the technical challenges of storage and transfer speeds of an ever-growing body of heterogeneous information; developing algorithms that can parse heterogeneous data with veracity so that downstream analyses are revealing; and designing analytical tools that can integrate molecular, clinical, demographic, and environmental elements that coalesce individual data points when the viewer steps away from the canvas. However, an over-arching challenge in this emerging field remains the development of tools to ensure the security of personal health information and scientific (e.g., genomic) data to maintain the privacy of patients. This challenge can be appreciated by considering the analogous problem of electronic fraud through identity theft, which is rampant in the developed, electronically-dependent world. These challenges notwithstanding, the ability to bring the power of vast amounts of electronic information to bear on the underpinnings of health and disease, the development of new pharmacological interventions, and the safety of the global formulary places biology and medicine on the verge of an exciting revolution.

Acknowledgments

SAW is the Samuel M.V. Hamilton Endowed Professor of Thomas Jefferson University. AT is Michael S. and Mary Sue Shannon Family Director, Center for Regenerative Medicine, and Marriott Family Professor of Cardiovascular Research at Mayo Clinic. This work was supported by grants from NIH (CA170533), Targeted Diagnostic & Therapeutics, Inc., and Mayo Clinic.

References

1. Waldman SA, Terzic A. Managing the innovation supply chain to maximize personalized medicine. Clin Pharmacol Ther. 2014; 95:113–8. [PubMed: 24448453]
2. Waldman SA, Terzic A. Molecular insights provide the critical path to disease mitigation. Clin Pharmacol Ther. 2014; 95:3–7. [PubMed: 24352148]
3. Wang H, Gu Q, Wei J, Cao Z, Liu Q. Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. Clin Pharmacol Ther. 2015; 97:451–4. [PubMed: 25670647]
4. Hersh W, et al. Health-care hit or miss? Nature. 2011; 470:327–9. [PubMed: 21331020]
5. Costa FF. Big data in biomedicine. Drug Discov Today. 2014; 19:433–40. [PubMed: 24183925]
6. Szlezak N, Evers M, Wang J, Perez L. The role of big data and advanced analytics in drug discovery, development, and commercialization. Clin Pharmacol Ther. 2014; 95:492–5. [PubMed: 24642713]
7. Schneweiss. (2016)
8. McDonagh E, Whirl-Carrillo M, Altman RB, Klein TE. Enabling the curation of your pharmacogenetic study. Clin Pharmacol Ther. 2015; 97:116–9. [PubMed: 25670512]
9. Roden. (2016)
10. Rasmussen-Torvik LJ, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther. 2014; 96:482–9. [PubMed: 24960519]
11. Van Driest SL, et al. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. Clin Pharmacol Ther. 2014; 95:423–31. [PubMed: 24253661]
12. Alfirevic A, et al. Phenotype standardization for statin-induced myotoxicity. Clin Pharmacol Ther. 2014; 96:470–6. [PubMed: 24897241]
13. Butte. (2016)
14. Harpaz. (2016)
15. Fang H, et al. Exploring the FDA adverse event reporting system to generate hypotheses for monitoring of disease characteristics. Clin Pharmacol Ther. 2014; 95:496–8. [PubMed: 24448476]
16. Pan D. 2016
17. Sarntivijai S, Abernethy DR. Use of internet search logs to evaluate potential drug adverse events. Clin Pharmacol Ther. 2014; 96:149–50. [PubMed: 25056395]
18. Shuren J. The FDA's role in the development of medical mobile applications. Clin Pharmacol Ther. 2014; 95:485–8. [PubMed: 24747239]