Big data in biomedicine

Fabricio F. Costa[1,2,3,4]

[1] Genomic Enterprise, Chicago, IL 60614, USA
[2] DataGenno Interactive Research Ltda, Rua Gastão Machado 66, Edifício CME, Salas 503/504, Campos dos Goytacazes, Rio de Janeiro, RJ 28035-120, Brazil
[3] 1871: DataGenno Interactive Research, 222 W. Merchandise Mart Plaza, 12th Floor, Suite 1212, Chicago, IL 60654, USA
[4] Cancer Biology and Epigenomics Program, Ann & Robert H. Lurie Children's Hospital of Chicago Research Center and Department of Pediatrics, Northwestern University's Feinberg School of Medicine, Chicago, IL 60614, USA

The increasing availability and growth rate of biomedical information, also known as 'big data', provides an opportunity for future personalized medicine programs that will significantly improve patient care. Recent advances in information technology (IT) applied to biomedicine are changing the landscape of privacy and personal information, with patients getting more control of their health information. Conceivably, big data analytics is already impacting health decisions and patient care; however, specific challenges need to be addressed to integrate current discoveries into medical practice. In this article, I will discuss the major breakthroughs achieved in combining omics and clinical health data in terms of their application to personalized medicine. I will also review the challenges associated with using big data in biomedicine and translational science

Introduction

A series of breakthroughs in medical science and IT are triggering a convergence between the healthcare industry and the life sciences industry that will quickly lead to more intimate and interactive relations among patients, their doctors and biopharmaceutical companies [1]. Big data analytics has an indispensable role in fostering those enhanced relations because it vastly enriches the remarkable but isolated wonder of the genome-on-a-thumb drive. Healthcare providers and drug makers now have the ability to explore and analyze omics data not only for an individual, but also in an aggregate from an increasing number of patients in specific population studies [2].

With rapid improvements in computer power, the cost of genome sequencing has plunged from millions of dollars per genome to thousands of dollars (and the cost will keep dropping). With advances in technology, patients will see a shift from population-based healthcare to personalized medicine that includes targeted diagnostics and treatment based on each patient's history, ancestry and genetic profile [3]. Online tools, such as the General Practice Research Database (GPRD), applied to clinical studies in drug discovery and assessment exemplify how IT is impacting biomedicine [4]. Although complex, this trend could one day revolutionize life sciences, biomedicine and what it means to be a healthcare professional or a researcher. The main benefits of applying big data analytics in personalized medicine include saving time while improving the overall quality and efficacy of treating disease.

Big data in biomedicine is driven by the single premise of one day having personalized medicine programs that will significantly improve patient care. Constant advances in understanding of different omics information are providing the footholds into establishing, for the first time, the causal genetic factors that could help manage the golden triangle of treatment: the right target, the right chemistry and the right patient. Solutions to deal with this overload of information are becoming a reality. However, challenges ahead include funneling clinical data, omics data, administrative data and also financial information securely into an unified system [5] to achieve better patient outcomes, advance research and continually improve the quality of patient care while reducing costs.

E-mail addresses: fcosta@luriechildrens.org, fcosta@genomicenteprise.com, fcosta@datagenno.com.

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.10.012

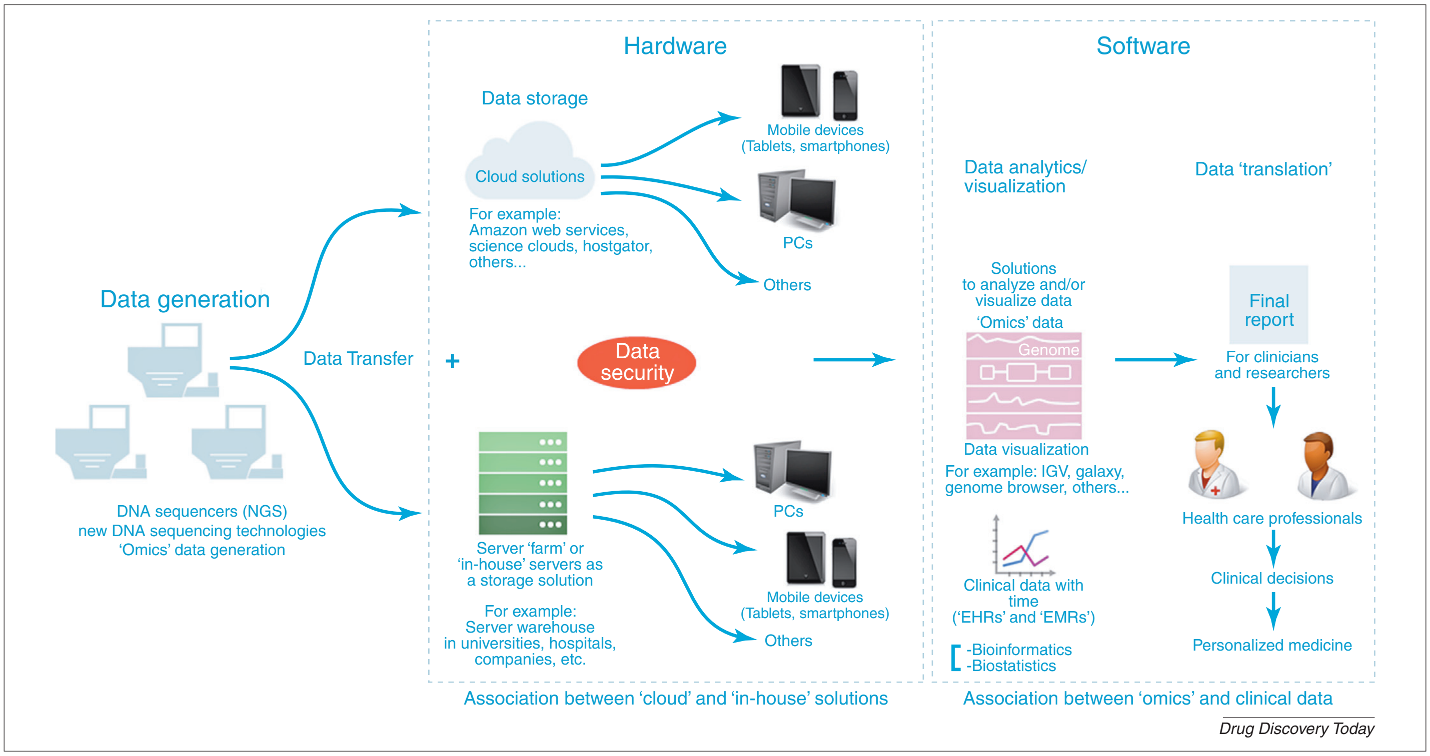

FIGURE 1

Big data in biomedicine. Schematic representation and depiction of a pipeline starting with data produced using next-generation sequencing (NGS), to data 'translation', and the generation of a 'final report' for clinicians and researchers. Personal health information and data generated by next-generation DNA sequencers (i.e. omics data such as genomes, transcriptomes, exomes, epigenomes and other types of similar information) are correlated, transferred to the 'cloud' or internal servers, analyzed and visualized using different solutions and tools that are available for big data analytics. Finally, data is translated as a short report to clinicians and researchers after a deep analysis for biomarkers and drug targets associated with specific disease phenotypes and after comparisons with public or private databases. Genome variants could be identified when comparing different samples, thus generating high-quality interpretation based on current knowledge and literature. This type of pipeline will ease the implementation and application of personalized medicine for clinicians and for research purposes. Between data transfer, storage and visualization, patient data needs to be secured by encryption of the information. Some solutions for both medical and scientific data security have been developed recently, but since this is a new area of study in biomedical informatics, big challenges lie ahead creating increasing opportunities in the market. Abbreviations: PCs, personal computers; EHRs, electronic health records; EMRs, electronic medical records; IGV, integrated genome viewer. Image designed by Eduardo Braga Ferreira Junior.

Although both computers and the internet have become faster, there is a lack of computational infrastructure that is needed to generate, maintain, transfer and analyze large-scale information securely in biomedicine and to integrate omics data with other data sets, such as clinical data from patients (Fig. 1). Indeed, it might now be less expensive to generate the data than it is to analyze and store it [6]. Another challenge is to transfer data from one location to another, because it is mainly done by mailing external drives with the information inside [6]. The security and privacy of the data from individuals are also a concern before and during data transfer [6]. Possible solutions to these issues include the use of better security systems with advanced encryption and de-identification algorithms, such as those used by banks in the financial sector to secure their clients' privacy [6]. The future of big data in life sciences is full of insecurities and challenges, but changes in several sectors are occurring to deal with it. Importantly, making sense of accumulating data in life sciences requires improved computational infrastructure, new methods to interpret the information and unique collaborative approaches.

In this article, I will discuss some of the major improvements in combining omics and clinical health data applied to personalized medicine. Moreover, an overview of the challenges faced by Big Data generation, transfer and analytics will be addressed. This article will also exemplify some of the major improvements needed to bridge the current technological gaps to address these challenges. Computational strategies, instrumentation and the current knowledge to interpret Big Data in order to make clinical decisions with a positive impact in biomedicine will also be discussed.

Big data, big impacts

Big data describe a new generation of technologies and architectures, designed to extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery and analysis [7]. This world of big data requires a shift in computing architecture so that researchers can handle both the data storage requirements and the heavy server processing needed to analyze large volumes of data in a secure manner [8]. Most of the big data surge is unstructured information and is not typically easy for traditional databases to analyze it. Therefore, the predictive power of big data has been explored recently in fields such as public health, science and medicine.

Computer tools to collect knowledge and insights from the vast trove of unstructured data available via the Internet are improving