Big Data Analytics and Cancer

Keywords: Big data; Biobank; Cloud computing; Cancer; Electronic Medical Records; Genomics; Proteogenomics

Opinion

The term big data has become a routine word across many disciplines [1-7]. The big data in medical terms generally encompasses Next Generation Sequencing (NGS) of the genome from individual patients, mRNA expression landscape of normal and diseased tissues, biobank tissue-derived information, clinical trials, drug efficacy and toxicology data and electronic medical records linked to medical imaging and insurance claims data [8-14]. During his State of the Union address (January 12, 2016), President Barack Obama announced the establishment of a Cancer Moonshot initiative to accelerate cancer research. This initiative, led by Vice President Joe Biden, aims to make therapies available to a large number of cancer patients and is projected to improve cancer prevention and detection at an early stage.

Recently (May 2016), the White House released The Federal Big Data Research and Development Strategic Plan, which provide guidance for developing or expanding Federal Big Data research and development (R&D) plans.

The Accelerating Medicines Partnership (AMP), a new venture involving the US National Institutes of Health (NIH), 10 biopharmaceutical companies, and several nonprofit organizations, has an initial fund of $230 Million. The overall goals are to transform the current approaches for diagnostics and treatments to a new dimension using big data analytics by jointly identifying and validating promising biological targets of disease. The initial therapeutic areas include Alzheimer’s disease, Type 2 diabetes and two autoimmune disorders, rheumatoid arthritis and systemic lupus erythematosus (lupus). The European drug research consortium projects that they will invest more than $5 billion in the next several years to apply big data techniques termed “Big Data for Better Outcomes,” to speed up clinical drug trials while developing a sustainable healthcare delivery system.

In the UK, the National Institute for Health Research (NIHR) has put in place a series of initiatives to help exploit the nation’s strengths in technology, medical research and healthcare data. The Genomics England Project is expected to generate a vast amount of genetic information from 100,000 patients with an initial focus on cancer, rare diseases and infectious diseases.

Among numerous therapeutic areas, cancer research area has accumulated huge amounts of big data [15-18]. This includes datasets from thousands of patients encompassing gene expression, mutations, deletions and amplifications and proteogenomic data [19-22]. Increasingly, the basic research in cancer is integrated into translational medicine in an attempt to move the discoveries closer to the clinic [13,23-25].

Key cancer-related big datasets include

The Cancer Genome Atlas (TCGA) research network: In collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), TCGA has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset to date incorporates 2.5 petabytes of data from tumor and matched normal tissues from more than 11,000 patients, is publically available [26];

The International Cancer Genome Consortium (ICGC): The ICGC data (release 22, Aug 2016) in total comprises data from more than 19,290 cancer donors spanning 70 projects and 21 tumor sites. The entire dataset is securely available on the Amazon Web Services (AWS) Cloud for access by cancer researchers worldwide [27];

Cancer Genome Hub at the University of California, Santa Cruz–UCSC: The Cancer Genomics Hub was established in August 2011 to provide a repository to TCGA. The CGHub has grown to be the largest database of cancer genomes in the world, storing more than 2.5 petabytes of data and serving downloads of nearly 3 petabytes per month [28];

The Catalogue of Somatic Mutations in Cancer (COSMIC): The COSMIC database is the world’s largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. The latest release (v70; Aug 2014) describes 2,002, 811 coding point mutations in over one million tumor samples and across most human genes [29];

The integrated cancer knowledgebase (canSAR): The canSAR database applies machine-learning approaches to provide drug-discovery predictions. The growing database now holds the 3D structures of almost three million cavities on the surface of nearly 110,000 molecules [30] and

The National Cancer Institute’s Clinical Proteomic Technologies for Cancer initiative: This database leverages proteogenomics analysis through the development of the Clinical Proteomic Tumor Analysis Consortium [31]. This consortium is composed of Proteome Characterization Centers, Data Center,
and Resources Center, to produce a unique continuum that defines the proteins translated from cancer genomes [22]. This integrative approach provides the broad scientific community with knowledge that links genotype to phenotype and ultimately phenotype. The data sets, analytically validated assays, as well as high quality reagents are publicly accessible. These efforts together with other NCI programs; e.g., the NCI’s Cancer Therapy Evaluation Program (CTEP), the Early Detection Research Network (EDRN), the Cooperative Groups have broadened the scope of cancer research from the bench to bedside.

Other cancer-related metadata includes the Oncomine® Gene Browser (ThermoFisher Scientific) dataset which harbors comprehensive gene profiles across thousands of cancer patient genomes with >500 sources [33]. The eBioPortal for cancer genomics which provides visualization, analysis and download of large-scale cancer genomics datasets [34] US Food and Drug Administration’s Mini-Sentinel [35], the National Patient-Centered Clinical Research Network- PCORNet [36], Claims datasets [37] and the American Society of Clinical Oncology’s CancerLinq [38].

Cloud-based computing efforts have greatly expanded the scope of mining the big data in cancer research by small to mid size research laboratories. The 1000 genomes Project cataloguing human sequence variations through deep sequencing of the 1000 genomes worldwide [39] uses a 200TB Amazon cloud-based data repository solution [40]. The Globus Genomics Systems [41] an Amazon cloud-based analysis and data management client is based on the open source, web-based Galaxy platform [42]. This system provides elastic scaling computer cluster infrastructure. Other data management systems that allow users to integrate large-scale genomics datasets include TranSMART [43], BioMart [44] and the Integrated Rule-oriented Data System (IRODS); open source data management software used by research organizations and government agencies worldwide. Google, Microsoft, Oracle and IBM also provide commercial cloud storage solutions used by research institutes including the National Institute of Health and the European Bioinformatics Institute.

In the area of breast cancer, the big data driven genomics has generated numerous “cancer signatures” which are being adopted into standard practice [22] such as the OncoType DX [45] and Mammaprint [46-48]. The “big data” analytics has also been used recently to predict if a patient is suffering from aggressive triple-negative breast cancer, slower-moving cancers or non-cancerous lesions with 95 percent accuracy [49].

Challenges

Significant challenges exist before the revolution in big data analytics can indeed benefit the vast number of cancer patients [50-52]. Both the basic researchers and practicing oncologists increasingly face the complexity of a plethora of bioinformatics tools and soft wares. Harnessing terabytes to exabytes of data emerging from numerous studies is a daunting task. Systems standardization across multiple platforms for the diverse tools needs to be established. The quality of datasets, the verification of tissue integrity and the electronic medical records are some of the areas requiring considerable improvements.

The soft wares used in the Electronic Medical Records (EMRs) are in a state of development. Integration of EMR with genomics data from individual patients faces considerable challenges. The GWAS big datasets encompass millions of single nucleotide variations (SNPs) amounting to terabytes of information [53,54]. Meaningful interpretations from these vast amounts of genetic data are difficult. Multiple platforms are being used to store the medical information, which are often not compatible [55-58]. This introduces a considerable level of complexity in deriving patient-centric information. Standards need to be introduced for the software used for the EMR.

The Ethical, Legal, and Social Implications (ELSI) of the worldwide genome initiatives continue to raise strong concerns [59]. Identification of fifty individuals from the 1000 genome project and public genealogy information using short tandem repeats [60], underscores this point. Together with the increasing use of cloud-based storage of the genomics data including the GWAS data, which matches genotypes to phenotypes, adds to the urgent need for clear guidelines to maintain privacy and security [61]. Development of de-identification algorithms [62,63] and customized user interface [64] could begin to address these concerns.

These issues notwithstanding, one can anticipate that the big data infrastructure should help the oncologists and cancer patients around the globe in decades to come. The big data cancer analytics with data encompassing clinical trials to real-world patients and practices can provide answers to effectiveness of treatment and long-term outcome.

References


