Opinion

Big data analytics and cancer

Keywords: big data, biobank, cloud computing, cancer, electronic medical records, genomics, proteogenomics

Volume 4 Issue 2 - 2016

Amy Makler, Ramaswamy Narayanan
Department of Biological Sciences, Florida Atlantic University, USA

Correspondence: Ramaswamy Narayanan, Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA, Tel +15612972247, Fax +15612973859, Email rnarayan@fau.edu

Received: October 03, 2016 | **Published:** October 13, 2016

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 it 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 $5billion 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 proteogenomics 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 2.5petabytes 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 2.5petabytes of data and serving downloads of nearly 3petabytes 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.[32] This integrative approach provides the broad scientific community with knowledge that links genotype to proteotype 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 cBioPortal 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 softwares. 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 softwares 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.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Costa FF. Big data in biomedicine. *Drug discovery today*. 2014;19(4):433–440.
2. Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*. 2016;5:12.
3. Xue LC, Dobbs D, Bonvin AM, et al. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett*. 2015;589(23):3516–3526.
4. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin Ther*. 2016;38(4):688–701.
5. Luo J, Wu M, Gopukumar D, et al. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights*. 2016;8:1–10.
6. Rein R, Memmert D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springerplus*. 2016;5(1):1410.
7. La Salle J, Williams KJ, Moritz C. Biodiversity analysis in the digital era. *Philos Trans R Soc Lond B Biol Sci*. 2016;371(1702):20150337.
8. Iorio F, Knijnenburg Theo A, Vis Daniel J, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. 2016;166(3):740–754.
9. Ciriello G, Miller ML, Aksoy BA, et al. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*. 2013;45(10):1127–1133.
10. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603–607.
11. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*. 2014;32(12):1202–1212.
12. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502(7471):333–339.
13. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC Health Serv Res*. 2010;10:231.
14. Peng H, Zhou J, Zhou Z, et al. Bioimage Informatics for Big Data. *Adv Anat Embryol Cell Biol*. 2016;219:263–272.
15. Coates J, Souhami L, El Naqa I. Big Data Analytics for Prostate Radiotherapy. *Front Oncol*. 2016;6:149.
16. Swift SL, Stojdl DF. Big Data Offers Novel Insights for Oncolytic Virus Immunotherapy. *Viruses*. 2016;8(2):E45.
17. Yang Y, Dong X, Xie B, et al. Databases and web tools for cancer genomics study. *Genomics Proteomics Bioinformatics*. 2015;13(1):46–50.

18. Kim ES. The Future of Molecular Medicine: Biomarkers, BATTLEs, and Big Data. *Am Soc Clin Oncol Educ Book*. 2015:22–27.

19. Chelala C, Hahn SA, Whiteman HJ, et al. Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. *BMC genomics*. 2007;8:439.

20. Barrett JH, Iles MM, Harland M, et al. Genome–wide association study identifies three new melanoma susceptibility loci. *Nat Genet*. 2011;43(11):1108–1113.

21. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–337.

22. Dawson SJ, Rueda OM, Aparicio S, Caldas C A new genome–driven integrated classification of breast cancer and its implications. *The EMBO journal*. 2013;32(5):617–628.

23. Meyer AM, Basch E. Big data infrastructure for cancer outcomes research: implications for the practicing oncologist. *J Oncol Pract*. 2015;11(3):207–208.

24. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. 2016;166(3):740–754.

25. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. *Clinical pharmacology and therapeutics*. 2016;99(3):285–297.

26. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan–Cancer analysis project. *Nature Genet*. 2013;45(10):1113–1120.

27. International Cancer Genome C, Hudson TJ, Anderson W, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993–998.

28. Cline MS, Craft B, Swatloski T, et al. Exploring TCGA Pan–Cancer Data at the UCSC Cancer Genomics Browser. *Scientific Reports*. 2013;3:2652.

29. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43(Database issue):D805–D811.

30. Tym JE, Mitsopoulos C, Coker EA, et al. canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res*. 2016;44(D1):D938–D43.

31. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*. 2014;513(7518):382–387.

32. Whiteaker JR, Halusa GN, Hoofnagle AN, et al. CPTAC Assay Portal: a repository of targeted proteomic assays. *Nature Methods*. 2014;11(7):703–704.

33. Rhodes DR, Kalyana–Sundaram S, Mahavisno V, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*. 2007;9(2):166–180.

34. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401–404.

35. Platt R, Carnahan R. The U.S. Food and Drug Administration's Mini–Sentinel Program. *Pharmacoepidemiology and Drug Safety*. 2012;21:1–303.

36. Fleurence RL, Curtis LH, Califf RM, et al. Launching PCORnet, a national patient–centered clinical research network. *JAMIA*. 2014;21(4):578–582.

37. Porter J, Love D, Costello A, et al. All–Payer Claims Database Development Manual: Establishing a Foundation for Health Care Transparency and Informed Decision Making. *APCD Council and West Health Policy Center*. 2015;96:2397–1053.

38. Schilsky RL, Michels DL, Kearbey AH, et al. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. *J Clin Oncol*. 2014;32(22):2373–2379.

39. The Genomes Project C. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.

40. Clarke L, Zheng–Bradley X, Smith R, et al. The 1000 Genomes Project: data management and community access. *Nature methods*. 2012;9(5):459–462.

41. Madduri RK, Sulakhe D, Lacinski L, et al. Experiences Building Globus Genomics: A Next–Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. *Concurr Comput*. 2014;26(13):2266–2279.

42. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11(8):R86.

43. Athey BD, Braxenthaler M, Haas M, et al. tranSMART: An Open Source and Community–Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:6–8.

44. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database*. 2011;2011:bar049.

45. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node–negative, estrogen receptor–positive breast cancer. *J Clin Oncol*. 2006;24(23):3726–3734.

46. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530–536.

47. Jezequel P, Campone M, Gouraud W, et al. bc–GenExMiner: an easy–to–use online platform for gene prognostic analyses in breast cancer. *Breast Cancer Res Treat*. 2012;131(3):765–775.

48. Hudis CA. Big data: Are large prospective randomized trials obsolete in the future? *Breast*. 2015;24 Suppl 2:S15–S18.

49. Agner SC, Rosen MA, Englander S, et al. Computerized Image Analysis for Identifying Triple–Negative Breast Cancers and Differentiating Them from Other Molecular Subtypes of Breast Cancer on Dynamic Contrast–enhanced MR Images: A Feasibility Study. *Radiology*. 2014;272(1):91–99.

50. Shaha SH, Sayeed Z, Anoushiravani AA, et al. Big Data, Big Problems: Incorporating Mission, Values, and Culture in Provider Affiliations. *Orthop Clin North Am*. 2016;47(4):725–732.

51. Chatellier G, Varlet V, Blachier–Poisson C, et al. "Big data" and "open data": What kind of access should researchers enjoy? *Therapie*. 2016;71(1):97–105..

52. Frelinger JA. Big Data, Big Opportunities, and Big Challenges. *J Investig Dermatol Symp Proc*. 2015;17(2):33–35.

53. Ramos EM, Hoffman D, Junkins HA, et al. Phenotype–Genotype Integrator (PheGenI): synthesizing genome–wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet*. 2014;22(1):144–147.

54. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res*. 2014;42(D1):D1001–D1006.

55. Farrugia G, Weinshilboum RM. Challenges in implementing genomic medicine: the Mayo Clinic Center for Individualized Medicine. *Clin Pharmacol Ther*. 2013;94(2):204–206.

56. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network:past, present, and future. *Genet Med*. 2013;15(10):761–771.

57. Kho AN, Rasmussen LV, Connolly JJ, et al. Practical challenges in integrating genomic data into the electronic health record. *Genet Med*. 2013;15(10):772–778.

58. Pathak J, Kho AN, Denny JC. Electronic health records–driven phenotyping: challenges, recent advances, and perspectives. *JAMIA*. 2013;20(e2):e206–e211.

59. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Sci Eng Ethics*. 2016;22(2):303–341.

60. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. *Science*. 2013;339(6117):321–324.

61. Skripcak T, Belka C, Bosch W, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. *Radiother Oncol*. 2014;113(3):303–309.

62. Schell SR. Creation of clinical research databases in the 21st century: a practical algorithm for HIPAA Compliance. *Surg Infect (Larchmt)*. 2006;7(1):37–44.

63. Fernandes AC, Cloete D, Broadbent MT, et al. Development and evaluation of a de–identification procedure for a case register sourced from mental health electronic records. *BMC Med Inform Decis Mak*. 2013;13:71.

64. Patel AA, Gilbertson JR, Showe LC, et al. A novel cross–disciplinary multi–institute approach to translational cancer research: lessons learned from Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC). *Cancer Inform*. 2007;3:255–274.